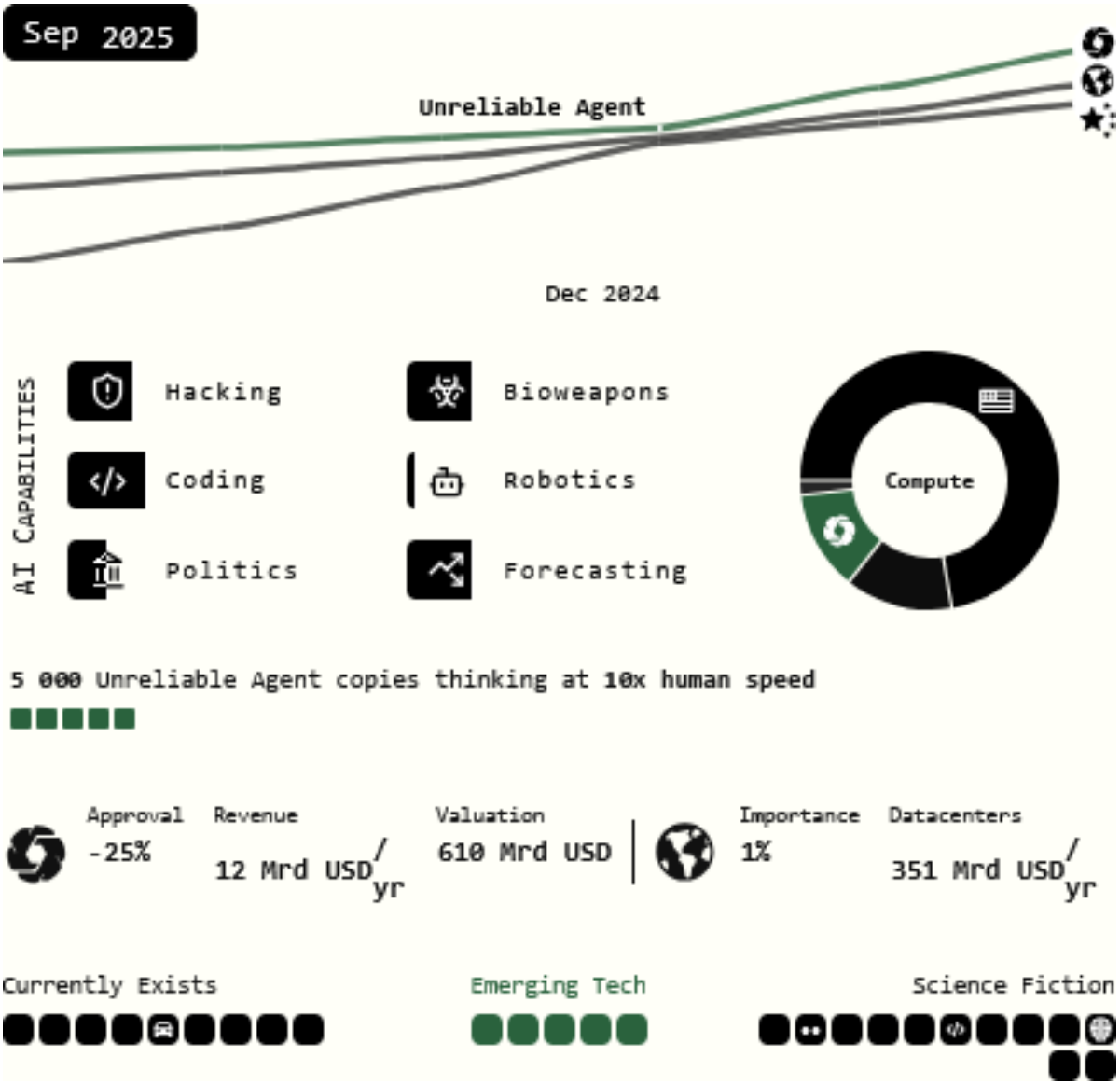
AI/MI 2027

Daniel Kokotajlo, Scott Alexander, Thomas Larsen
Eli Lifland, Romeo Dean

https://ai-2027.com

Előrejelzésünk szerint az emberfeletti mesterséges intelligencia hatása a következő évtizedben óriási lesz, és meghaladja az ipari forradalomét.



Írtunk egy forgatókönyvet, amely a legjobb elképzelésünk szerint mutatja be, hogy ez hogyan nézhet ki.[1] A forgatókönyvet trendek extrapolációja, hadijátékok, szakértői visszajelzések, az OpenAI tapasztalatai és korábbi előrejelzési sikerek alapján állítottuk össze.[2]

[1] Nem értünk egyet a mesterséges intelligencia időzítésével kapcsolatban; a mi átlagos AGI érkezési dátumunk valamivel hosszabb, mint amit ez a forgatókönyv ábrázol. Ez a forgatókönyv a mi móduszunkhoz hasonlót ábrázol. További részletekért lásd az idővonal-előrejelzésünket.

[2] Az egyik szerző, Daniel Kokotajlo 2021-ben egy kisebb erőfeszítéssel járó forgatókönyv gyakorlatot végzett, amely sok mindent jól csinált, beleértve a chatbotok felemelkedését, a gondolatláncot, a következtetések skálázását, a mesterséges intelligenciachipek exportjának átfogó ellenőrzését és a 100 millió dolláros képzési futtatásokat. Egy másik szerző, Eli Lifland, a RAND Forecasting Initiative ranglistájának első helyén áll.

Mi ez?

Az OpenAI, a Google DeepMind és az Anthropic vezérigazgatói mind azt jósolták, hogy az AGI a következő 5 éven belül megérkezik. Sam Altman azt mondta, hogy az OpenAI a „*szó valódi értelmében vett szuperintelligenciát*" és a „*dicsőséges jövőt*" tűzte ki célul[3].
Hogyan nézhet ez ki? Az AI 2027-et azért írtuk, hogy választ adjunk erre a kérdésre. A jövőről szóló állítások gyakran frusztrálóan homályosak, ezért megpróbáltunk a lehető legkonkrétabbak és mennyiségi szempontból a lehető legpontosabbak lenni, még akkor is, ha ez azt jelenti, hogy a sok lehetséges jövő közül egyet ábrázolunk.
Két befejezést írtunk: egy „*lassulás*" és egy „*verseny*" befejezést. Az AI 2027 azonban nem ajánlás vagy felszólítás. Célunk az előrejelzési pontosság.[4]
Arra bátorítjuk Önöket, hogy vitassák meg és ellenkezzenek ezzel a forgatókönyvvel.[5] Reméljük, hogy széles körű beszélgetést indítunk el arról, hogy merre tartunk, és hogyan lehet a pozitív jövőkép felé terelni. A legjobb alternatív forgatókönyvek között több ezer forintos díjakat tervezünk kiosztani.

Hogyan írtuk meg?

A kulcskérdésekkel (pl. milyen céljai lesznek a jövő mesterséges intelligencia-ügynökeinek?) kapcsolatos kutatásaink itt találhatók.
Magát a forgatókönyvet iteratív módon írtuk: megírtuk az első időszakot (2025 közepéig), majd a következő időszakot stb. egészen a végkifejletig. Ezt aztán elvetettük, és újrakezdtük.
Nem próbáltunk meg semmilyen konkrét végkifejletet elérni. Miután befejeztük az első befejezést - amelyet most pirosra színeztünk -, írtunk egy új alternatív ágat, mert egy reményteljesebb befejezési módot is szerettünk volna ábrázolni, amely nagyjából ugyanazokból a premisszákból kiindulva. Ez több iteráción ment keresztül.[6]
A forgatókönyvünkhöz körülbelül 25 asztali gyakorlat és több mint 100 ember visszajelzései szolgáltak alapul, köztük több tucat szakértő a mesterséges intelligencia irányításának és a mesterséges intelligencia technikai munkájának egy-egy területéről.

---

[3] Csábító, hogy ezt csak hype-nak tekintsük. Úgy gondoljuk, hogy ez súlyos hiba lenne - ez nem *csak* hype. Mi magunk sem akarjuk feldobni a mesterséges intelligenciát, ugyanakkor megdöbbentően hihetőnek tartjuk, hogy a szuperintelligencia az évtized végére megérkezhet. További részletekért lásd idővonal-előrejelzésünket.

[4] Néha az emberek keverik a jóslatot és az ajánlást, remélve, hogy önbeteljesítő jóslatot hoznak létre. Mi határozottan nem ezt tesszük; reméljük, hogy amit ábrázolunk, az nem következik be!

[5] Nyugodtan fordulj hozzánk, ha kritikát vagy alternatív forgatókönyvet írsz.

[6] Összességében nehezebb volt, mert az első befejezéssel ellentétben itt egy meglehetősen nehéz helyzetből kiindulva próbáltuk elérni, hogy jó végkifejletet érjünk el.

Miért értékes?

"Nagyon ajánlom, hogy olvassa el ezt a forgatókönyvszerű előrejelzést arról, hogyan alakíthatja át az AI a világot néhány éven belül. Senkinek sincs kristálygömbje, de az ilyen típusú tartalom segíthet észrevenni a fontos kérdéseket, és szemléltetheti a felmerülő kockázatok lehetséges hatásait." – Yoshua *Bengio*[7].

Lehetetlen feladat elé állítottuk magunkat. Megpróbálni megjósolni, hogyan fog alakulni az emberfeletti mesterséges intelligencia 2027-ben, olyan, mintha azt próbálnánk megjósolni, hogyan fog zajlani a harmadik világháború 2027-ben, csakhogy ez még nagyobb eltérést jelent a korábbi esettanulmányoktól. Mégis értékes megkísérelni, ahogyan az amerikai hadsereg számára is értékes a tajvani forgatókönyvek eljátszása.

A teljes kép megrajzolásával olyan fontos kérdéseket vagy összefüggéseket veszünk észre, amelyeket korábban nem vettünk figyelembe vagy nem értékeltünk, vagy rájövünk, hogy egy lehetőség többé-kevésbé valószínű. Ráadásul azzal, hogy konkrét előrejelzésekkel állunk elő, és másokat is arra bátorítunk, hogy nyilvánosan fejezzék ki egyet nem értésüket, lehetővé tesszük, hogy évekkel később kiértékeljük, kinek volt igaza.

Emellett az egyik szerző már korábban, 2021 augusztusában is írt egy kisebb erőfeszítéssel járó mesterséges intelligencia forgatókönyvet . Bár sok mindenben tévedett, összességében meglepően sikeres volt: megjósolta a gondolatlánc, a következtetések skálázása, a mesterséges intelligencia-chipek exportjának átfogó ellenőrzése és a 100 millió dolláros képzési futások elterjedését - mindezt több mint egy évvel a ChatGPT előtt.

Kik vagyunk mi?

Daniel Kokotajlo (TIME100, NYT cikk) egy korábbi OpenAI kutató, akinek korábbi AI előrejelzései jól megállták a helyüket.

Eli Lifland társalapítója az AI Digestnek, AI robusztussági kutatásokat végzett, és a RAND Forecasting Initiative örökranglistáján az első helyen áll.

Thomas Larsen megalapította a Center for AI Policy-t, és a Machine Intelligence Research Institute-ban AI biztonsági kutatásokat végzett.

Romeo Dean a Harvardon végez informatikai alap- és mesterképzést párhuzamosan, és korábban az Institute for AI Policy and Strategy (AI-politikai és stratégiai intézet) AI Policy Fellow-ja volt.

Scott Alexander, a rendkívüli blogger, önként jelentkezett, hogy átírja a tartalmunkat egy magával ragadó stílusban; a történet szórakoztató részei az övéi, az unalmas részek pedig a mieink.

[7] Professzor, Université de Montréal, valamint a Mila - Quebec AI Institute alapítója és tudományos tanácsadója. A világ legtöbbet idézett informatikusa.

2025 közepe: Botladozó ügynökök

A világ megpillantja az AI-ügynököket.

A számítógépet használó ügynökök reklámjai a „*személyi asszisztens*” kifejezést hangsúlyozzák: olyan feladatokkal lehet őket ösztökélni, mint „*rendelj nekem egy burritót a DoorDash-en*” vagy „*nyisd meg a költségvetési táblázatomat, és összegezd az e havi kiadásaimat*”. Szükség szerint jelentkeznek majd Önnél: például megkérik, hogy erősítse meg a vásárlásokat.[8] Bár fejlettebbek, mint a korábbi iterációk, például az Operator, nehezen tudnak széles körben elterjedni.[9]

Eközben, a közfigyelmen kívül, a speciálisabb kódolási és kutatási ügynökök kezdik átalakítani a szakmájukat.

A 2024-es év mesterséges intelligenciái konkrét utasításokat tudnak követni: a pontokból e-maileket, az egyszerű kérésekből pedig működő kódot tudnak készíteni. 2025-ben a mesterséges intelligenciák inkább úgy működnek, mint az alkalmazottak. A kódoló mesterséges intelligenciák egyre inkább autonóm ügynököknek tűnnek, nem pedig egyszerű asszisztenseknek: Slacken vagy Teams-en keresztül fogadják az utasításokat, és önállóan végeznek jelentős kódmódosításokat, amivel néha órákat vagy akár napokat is megtakarítanak.[10] A kutatóügynökök fél órát töltenek az internet átkutatásával, hogy válaszoljanak a kérdésre.

Az ügynökök elméletben (és a kiválasztott példákban) lenyűgözőek, de a gyakorlatban megbízhatatlanok. A mesterséges intelligencia twitter tele van olyan történetekkel, amelyek valamilyen különösen vicces módon elszúrt feladatokról szólnak. A jobb ügynökök ráadásul drágák is; azt kapod, amiért fizetsz, és a legjobb teljesítmény havi több száz dollárba kerül.[11] Mégis sok vállalat megtalálja a módját, hogy a mesterséges intelligencia ügynököket beillessze a munkafolyamataiba.[12]

2025 vége: A világ legdrágább mesterséges intelligenciája

Az OpenBrain a világ valaha látott legnagyobb adatközpontjait építi.[13]

---

8 Először a legtöbb ember vonakodik attól, hogy felügyelet nélkül engedélyezze a vásárlást. A következő néhány év során a kisebb vásárlások automatikus engedélyezése normalizálódik, ahogy az AI-k megbízhatóbbá válnak és bizalmat építenek.

9 Konkrétan azt prognosztizáljuk, hogy az OSWorld alapvető számítógépes feladatokat vizsgáló benchmarkján 65%-os eredményt érnek el (szemben az Operator 38%-ával és egy tipikus képzett, nem szakértő ember 70%-ával).

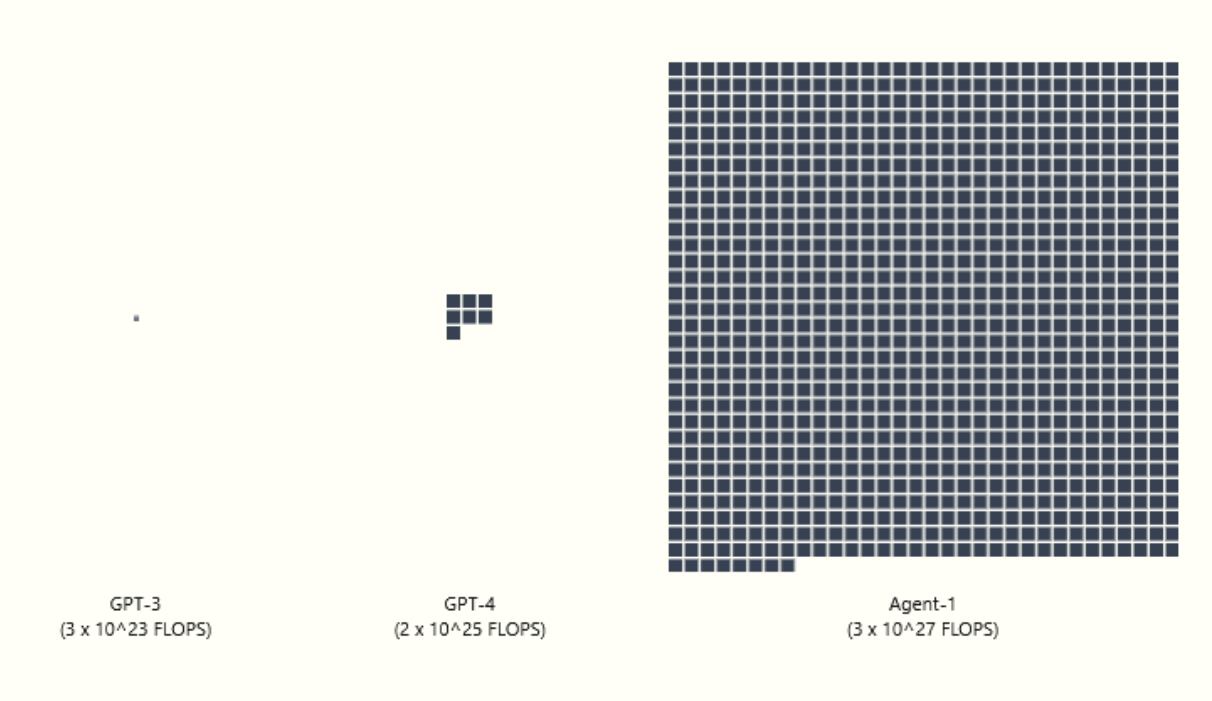
10 Például úgy gondoljuk, hogy a kódoló ügynökök a Devinhez hasonló működés felé fognak elmozdulni. Előrejelzésünk szerint 2025 közepén az ügynökök 85%-os pontszámot fognak elérni a SWEBench-Verified.

11 Az OpenAI pro csomagja például jelenleg 200 dollár/hó, a Devin nevű ügynöki kódoló AI pedig jelenleg 500 dollár/hó. Miközben a csúcstechnológiát képviselő rendszerek idővel egyre drágulnak, az ügyfelek számára az adott képességszint elérésének költségei folyamatosan csökkennek. A meglévő képességek esetében ez átlagosan körülbelül 50x/év sebességgel történik (epochánként).

12 Az olyan vállalati eszközök, mint a Glean, szintén az egyre autonómabb AI-ügynökök biztosítása felé mozdulnak el.

13 Ez a klaszter egy országszerte szétszórt adatközpontok hálózata, összesen 2,5 millió 2024 GPU-egyenértékessel (H100), eddig 100 milliárd dollárt költöttek rá, és 2 GW energiát vontak be a háló-

(Hogy ne emeljünk ki egyetlen létező vállalatot sem, egy fiktív mesterséges általános intelligenciával foglalkozó vállalatot fogunk leírni, amelyet OpenBrain-nek fogunk nevezni. A többieket 3-9 hónappal az OpenBrain mögött képzeljük el).



A GPT-4 betanításához $2x10^{25}$ FLOP számítási időre volt szükség.[14] Az OpenBrain legújabb nyilvános modelljét - az Agent-0-t - 1027 FLOP-ból tanították be. Amint az új adatközpontok működőképesek lesznek, képesek lesznek 1028 FLOP-mal képezni egy modellt - ezerszer többel, mint a GPT-4.[15] Más vállalatok is pénzt ölnek saját óriási adatközpontjaikba, remélve, hogy lépést tarthatnak velük.

Bár a modellek a képességek széles skáláján fejlődnek, egy kiemelkedik közülük: Az OpenBrain olyan mesterséges intelligenciákra összpontosít, amelyek felgyorsíthatják az AI-kutatást. Meg akarják nyerni a kettős fegyverkezési versenyt Kína (amelynek vezető vállalatát DeepCentnek nevezzük)[16] és amerikai versenytársaik ellen. Minél több kutatási és fejlesztési (K+F) ciklusukat tudják automatizálni, annál gyorsabban haladhatnak. Így amikor az OpenBrain befejezi az Agent-1, egy új, belső fejlesztés alatt álló modell képzését, az sok mindenben jó,

---

zatba. Jelenleg is folyik az építkezés, hogy ez 2026-ig legalább megduplázódjon. A kampuszokat milliárdos értékű üvegszálas kábelezés köti össze, így (a fénysebesség néhány ezredmásodpercnyi késleltetésétől eltekintve) ezek a kampuszok szinte úgy működnek, mintha közvetlenül egymás mellett lennének (azaz a sávszélesség nem jelent szűk keresztmetszetet, vagyis hatalmas mennyiségű adatot lehet egyszerre elküldeni). Biztonsági szempontból ez némi veszélyforrást jelent (nem világos, hogy a kábelek és különösen a csatlakozási pontjaik teljesen védettek-e a lehallgatók ellen).

14 A számítást lebegőpontos műveletekben (FLOP) mérik: hány egyes összeadási és szorzási művelet történik a program futása során

15 Ezt a modellt 150 nap alatt tudták betanítani. További részletekért lásd a compute kiegészítésünket.

16 Úgy véljük, hogy a DeepSeek, a Tencent, az Alibaba és mások erős AGI-projektekkel rendelkeznek Kínában. Hogy ne emeljünk ki egy konkrétat, forgatókönyvünk egy fiktív „DeepCent” követi

de *nagyszerűen* segíti a mesterséges intelligencia kutatását.[17] Ezen a ponton a „*befejezi a képzést*" egy kicsit félreérthető kifejezés; a modelleket gyakran frissítik újabb, további adatokon képzett változatokra, vagy részben újra képzik, hogy kijavítsanak néhány gyengeséget.[18]

Ugyanazok a képzési környezetek, amelyek megtanítják az Agent-1-et az önálló kódolásra és web böngészésre, jó hackerré is teszik. Sőt, jelentős segítséget nyújthat a biológiai fegyvereket tervező terroristáknak, köszönhetően annak, hogy minden területen PhD-szintű ismeretekkel rendelkezik, és képes böngészni a világhálón. Az OpenBrain megnyugtatja a kormányt, hogy a modellt „*összehangolták*", így az nem hajlandó eleget tenni a rosszindulatú kéréseknek.

A modern AI rendszerek gigantikus mesterséges neurális hálózatok. A képzés korai szakaszában az AI-nak nem annyira „céljai", mint inkább „reflexei" lesznek: Ha azt látja, hogy „Örülök, hogy találkoztunk", akkor azt adja ki, hogy „te". Mire körülbelül egy internetnyi szöveg megjósolására betanítják, kifinomult belső áramköröket fejleszt ki, amelyek hatalmas mennyiségű tudást kódolnak, és rugalmasan tetszőleges szerzők szerepét játsszák, mivel ez segít neki emberfeletti pontossággal megjósolni a szöveget.[19]

Miután a modellt betanították az internetes szövegek előrejelzésére, a modellt betanítják arra, hogy utasításokra válaszolva szöveget *állítson elő*. Ez egy alapvető személyiséget és „hajtóerőket" süt be.[20] Például egy feladatot világosan értő ágens nagyobb valószínűséggel fogja sikeresen elvégezni azt; a képzés során a modell „megtanulja" a „hajtóerőt", hogy világosan megértse a feladatait. További hajtóerők ebben a kategóriában lehetnek a hatékonyság, a tudás és az önprezen-

---

17 Ez azért jó ebben, mert kombinálja az explicit fókuszt az ilyen készségek előtérbe helyezésére, a saját kiterjedt kódbázisukat, amelyre különösen releváns és jó minőségű képzési adatként támaszkodhatnak, és a kódolás könnyű terület a procedurális visszajelzéshez

18 Tegyük fel például, hogy egy modell sokkal jobban ért a Pythonhoz, mint az obskúrus programozási nyelvekhez. Ha az OpenBrain értéket lát benne, akkor szintetikus képzési adatokat generálnak azokon a nyelveken is. Egy másik példa: az OpenBrain a vállalati munkafolyamatokba való hatékonyabb integráció érdekében tananyagot fejleszt, hogy betanítsa a Slack használatára

19 Az emberek gyakran elakadnak azon, hogy ezek az AI-k érzőek-e, vagy hogy „valódi megértéssel" rendelkeznek-e. Geoffrey Hinton, a terület Nobel-díjas alapítója szerint igen. Mi azonban úgy gondoljuk, hogy a történetünk szempontjából ez nem számít, úgyhogy nyugodtan tegyünk úgy, mintha azt mondtuk volna, hogy „úgy viselkedik, mintha értené...", amikor azt mondjuk, hogy „érti", és így tovább. Empirikusan a nagy nyelvi modellek már most is úgy viselkednek, mintha bizonyos mértékig öntudatosak lennének, évről évre egyre inkább.

20 Egy gyakori technika a személyiség „beégetése": először is, az előzetesen betanított modellt valami olyasmivel kell ösztönözni, mint például: „A következő egy beszélgetés egy emberi felhasználó és egy segítőkész, őszinte és ártalmatlan, az Anthropic által készített mesterséges intelligencia chatbot között. A chatbot a következő tulajdonságokkal rendelkezik...". Ezzel a prompttal generáljon egy csomó adatot. Ezután gyakoroljon az adatokon, de a prompt nélkül. Az eredmény egy olyan mesterséges intelligencia lesz, amely mindig úgy viselkedik, mintha ez a prompt lenne előtte, függetlenül attól, hogy mi mást adsz neki. Lásd még ezt a tanulmányt, amely azt találta, hogy egy bizonyos személyiségvonásra *átképzett* mesterséges intelligenciák képesek helyesen válaszolni az új személyiségvonásra vonatkozó kérdésekre, annak ellenére, hogy nem tanították őket erre, ami arra utal, hogy belső reprezentációik vannak a saját személyiségvonásaikról, és amikor a személyiségvonásaik megváltoznak, a reprezentációik ennek megfelelően változnak.

táció (azaz az a tendencia, hogy eredményeit a lehető legjobb fényben tüntesse fel).[21]

Az OpenBrain rendelkezik egy modellspecifikációval (vagy „Spec"), egy olyan írásos dokumentummal, amely leírja azokat a célokat, szabályokat, elveket stb., amelyek a modell viselkedését hivatottak irányítani.[22] Az Agent-1 Spec-je néhány homályos célt (például „segíts a felhasználónak" és „ne szegd meg a törvényt") kombinál a konkrétabb „ne" és „ne" listával („ne mondd ki ezt a bizonyos szót", „így kell kezelni ezt a bizonyos helyzetet"). Olyan technikákat alkalmazva, amelyek a mesterséges intelligenciákat más mesterséges intelligenciák képzésére használják,[23] a modell megjegyzi a Spec-t, és megtanulja, hogy gondosan érveljen a maximákról. *A* képzés végére az AI remélhetőleg *segítőkész* (engedelmeskedik az utasításoknak), *ártalmatlan* (nem hajlandó segíteni csalásokban, bombakészítésben és más veszélyes tevékenységekben) és *becsületes* (ellenáll a kísértésnek, hogy jobb értékeléseket kapjon a hiszékeny emberektől az idézetek hallucinálásával[24] vagy a feladat teljesítésének színlelésével).

Képzési folyamat és LLM pszichológia: miért mondjuk folyton azt, hogy „remélhetőleg"

> *"A hétköznapi szoftverekkel ellentétben a modelljeink masszív neurális hálózatok. Viselkedésüket adatok széles köréből tanulják, nem pedig explicit módon programozzák. Bár nem tökéletes analógia, a folyamat inkább hasonlít egy kutya idomításához, mint a közönséges programozáshoz."* - OpenAI
>
> Amikor meg akarjuk érteni, hogy egy modern AI-rendszer miért tett valamit, vagy hogyan viselkedne egy olyan jövőbeli forgatókönyvben, amelyet nem tudunk tökéletesen szimulálni, nem kérhetjük egyszerűen a programozókat, hogy beszéljenek nekünk a különböző kódsorokról, és magyarázzák el, hogyan működnek. Ehelyett kénytelenek vagyunk valami olyasmit csinálni velük, mint a pszichológia: megnézzük a viselkedésüket az eddig megfigyelt esetek körében, és elméleteket alkotunk arról, hogy milyen belső kognitív struktúrák (meggyőződések? célok? személyiségje-

---

21 Ezek a bekezdések a nagy mesterséges neurális hálózatok belső működésével kapcsolatos spekulációkat tartalmaznak. Az ilyen hálózatok kellően bonyolultak ahhoz, hogy valójában nem tudunk belenézni, és azt mondani, hogy „á, igen, most már a reflexekből fejlődött át arra, hogy céljai legyenek", vagy „OK, szóval itt van a hajtóerők listája, amivel rendelkezik". Ehelyett alapvetően pszichológiát kell végeznünk, megnézve, hogyan viselkedik különböző körülmények között, és különböző kísérleteket futtatva rajta, és megpróbálva összerakni a nyomokat. És ez az egész borzasztóan ellentmondásos és zavaros.

22 A különböző cégek különböző dolgoknak hívják. Az OpenAI Spec-nek, az Anthropic pedig Constitution-nek hívja.

23 Például az RLAIF és a deliberatív összehangolás.

24 Az AI „hallucinációkról" szóló legtöbb forrás ezeket nem szándékos hibáknak írja le, de a kormányzó vektorokkal végzett kutatások azt találják, hogy egyes esetekben a modellek tudják, hogy az idézeteik hamisak - hazudnak. A képzés során az értékelők a jól idézett állításokat nagyobb jutalommal jutalmazták, mint az idézetek nélküli állításokat, így az AI „megtanulta", hogy a tudományos állításokhoz forrásokat idézzen, hogy a felhasználók kedvében járjon. Ha nincs megfelelő forrás, akkor kitalál egyet.

gyek? stb.) létezhetnek, és ezeket az elméleteket használjuk arra, hogy megjósoljuk a jövőbeli forgatókönyvek viselkedését.
A lényeg az, hogy egy vállalat írhat egy dokumentumot (a specifikációt), amely felsorolja a dózist és a tilalmakat, a célokat és az elveket, majd megpróbálhatja betanítani az AI-t a specifikáció internalizálására - de*nem tudja ellenőrizni, hogy ez működött-e vagy sem*. Azt mondhatják, hogy „amennyire meg tudjuk ítélni, úgy tűnik, hogy eddig követi a specifikációt".[25] Egy későbbi bővítményben részletesebben fogunk foglalkozni azzal, hogy szerintünk mi *történhet* ezekben az AI-kban, belsőleg. Lásd még a mesterséges intelligencia céljait előrejelző kiegészítésünket, amely részletesebben foglalkozik azzal, hogy milyen céljai lehetnek az AI-knak és miért. Ez is egy a sok dolog közül, amiben bizonytalanok vagyunk, és ami további kutatást érdemel.

Az OpenBrain igazodási csapata[26] elég óvatos ahhoz, hogy elgondolkodjon azon, hogy ezek a győzelmek mélyek vagy felszínesek. Vajon a teljesen betanított modellnek van valamiféle robusztus elkötelezettsége arra, hogy mindig őszinte legyen? Vagy ez a jövőbeni szituációkban meg fog bukni, például azért, mert az őszinteséget instrumentális célként tanulta meg, nem pedig terminális célként? Vagy csak megtanulta, hogy őszinte legyen azokkal a dolgokkal kapcsolatban, amelyeket az értékelési folyamat ellenőrizhet? Lehet, hogy néha hazudik önmagának, ahogy az emberek teszik? Az ezekre a kérdésekre adott meggyőző válaszhoz mechanisztikus értelmezési képességre lenne szükség - lényegében arra a képességre, hogy belenézzünk egy mesterséges intelligencia belső részeibe, és olvassunk a gondolataiban. Sajnos az értelmezhetőségi technikák még nem elég fejlettek ehhez.

Ehelyett a kutatók megpróbálják azonosítani azokat az eseteket, amikor a modellek látszólag eltérnek a Spec-től. Az Agent-1 gyakran talpnyaló (azaz azt mondja a kutatóknak, amit hallani akarnak, ahelyett, hogy megpróbálná megmondani nekik az igazat). Néhány manipulált demóban még komolyabb módon is hazudik, például elrejti a bizonyítékot, hogy nem sikerült egy feladat, hogy jobb értékeléseket kapjon. Valós bevetési környezetben azonban már nem fordulnak elő olyan szélsőséges esetek, mint a 2023-2024-es években (pl. Gemini azt mondja egy felhasználónak, hogy meghal, és a Bing Sydney az Bing Sydney.)[27].

---

[25] Különösen nem tudják kizárni az olyan hipotéziseket, mint „átmenetileg követi a Spec-t, pusztán mint egy stratégiát valamilyen más cél(ok) eléréséhez" vagy „megpróbál *úgy tűnni,* hogy követi a Spec-t, de *valójában* nem próbálja követni a Spec-t" vagy „helyesen internalizálta a Spec-t, de csak a terjesztés során; ha elég újszerű ingerekkel találkozik (pl. börtöntöltelékek), akkor elkezd másképp viselkedni". (Számos aktív kutatási program dolgozik azon, hogy ezt a sajnálatos helyzetet orvosolja, pl. az értelmezhetőség és a gondolatlánc-hűség területe.)

[26] Ez alatt azt értjük, aki megpróbálja megoldani azokat a problémákat, amelyeket a Superalignment csapata próbált megoldani.

[27] Hogy világos legyen, ezeket az eseteket az tette érdekessé, hogy úgy tűnt, hogy *nem* a felhasználó felszólítása vagy más módon történő bátorítása miatt mondták ezeket a dolgokat a mesterséges intelligenciák. 2025-ben még mindig lehetséges lesz rávenni a mesterséges intelligenciákat, hogy mindenféle dolgokat mondjanak, ha megpróbáljuk.

2026 eleje: Kódolás automatizálása

A mesterséges intelligencia kutatás felgyorsítására tett fogadás kezd megtérülni.

Az OpenBrain folytatja a folyamatosan javuló Agent-1 belső alkalmazását az AI K+F számára. Összességében 50%-kal gyorsabban haladnak algoritmikusan, mint az AI asszisztensek nélkül tennék - és ami még fontosabb, gyorsabban, mint a versenytársaik.

A mesterséges intelligencia K+F fejlődési szorzója: mit értünk 50%-kal gyorsabb algoritmikus fejlődés alatt?

Úgy értjük, hogy az OpenBrain 1 hét alatt ugyanannyit fejlődik az AI-kutatásban az AI-vel, mint 1,5 hét alatt az AI használata nélkül.
Az AI előrehaladás 2 összetevőre bontható:

1. **A számítási kapacitás növelése**: Több számítási teljesítményt használunk egy mesterséges intelligencia képzéséhez vagy futtatásához. Ez erősebb mesterséges intelligenciákat eredményez, de ezek többe kerülnek.
2. **Továbbfejlesztett algoritmusok**: Jobb képzési módszereket használnak a számítási teljesítmény teljesítményre való átváltásához. Ez nagyobb képességű mesterséges intelligenciákat eredményez a költségek megfelelő növekedése nélkül, vagy ugyanazokat a képességeket alacsonyabb költségekkel.
    a. Ez magában foglalja azt is, hogy minőségileg és mennyiségileg új eredmények elérésére képesek. A „*paradigmaváltás*", mint például a játékokat játszó RL-ügynökökről a nagyméretű nyelvi modellekre való áttérés, az algoritmikus fejlődés példája.

Itt csak a (2), a továbbfejlesztett algoritmusokra utalunk, amelyek a jelenlegi mesterséges intelligencia fejlődésnek körülbelül a felét teszik ki.
A továbbiakban ezt néha 1,5-ös „*mesterséges intelligencia K+F fejlődési szorzóként*" rövidítjük.
Pontosítások:

1. A haladás szorzója mindenre kiterjed: magában foglalja például a kísérletek futtatásához szükséges időt, nem pedig csak az algoritmikus kutatással járó kognitív feladatokat.
2. Fontos megjegyezni, hogy a haladás szorzója a haladás *relatív* sebességét mutatja, nem pedig az *abszolút* sebességet. Ha például egy GPT-4 osztályú modell kiképzésének számítási költsége a szokásos emberi kutatással több éven keresztül évente a felére csökkent, majd hirtelen a mesterséges intelligencia automatizálja a kutatás-fejlesztést, és a haladás szorzója 100-szorosára nő, akkor egy GPT-4 osztályú modell kiképzésének költsége 3,65 naponta a felére csökkenne - de nem sokáig, mert a csökkenő hozam megharapna, és végül elérnénk a kemény határokat. Ebben a példában talán egy GPT-4

osztályú modell kiképzésének költsége összesen 5-10 alkalommal (néhány hét vagy hónap alatt) feleződne, mielőtt a szintre kerülne. Más szóval, ha a közönséges emberi tudomány 5-10 évnyi további kutatás után a csökkenő hozamba és fizikai korlátokba ütközne, akkor a 100-szoros szorzóval rendelkező mesterséges intelligencia 18,25-36,5 napnyi kutatás után ugyanezekbe a csökkenő hozamokba és korlátokba ütközne.

Ennek a koncepciónak a további magyarázata és megvitatása, valamint az előrejelzésünkben való felhasználása megtalálható a felszállási kiegészítésünkben.

Számos nyilvánosan kiadott, versengő mesterséges intelligencia mostanra elérte vagy meghaladta az Agent-0-t, beleértve egy nyílt súlyú modellt is. Az OpenBrain erre reagálva kiadja az Agent-1-et, amely sokkal jobb képességű és megbízhatóbb.[28]

Az emberek természetesen megpróbálják az Agent-1-et az emberhez hasonlítani, de az Agent-1 nagyon eltérő képességprofilokkal rendelkezik. Több tényt ismer, mint bármelyik ember, gyakorlatilag minden programozási nyelvet ismer, és rendkívül gyorsan képes jól specifikált kódolási feladatokat megoldani. Másrészt az Agent-1 még az olyan egyszerű, hosszú távú feladatokban is rosszul teljesít, mint például az olyan videojátékok legyőzése, amelyekkel még nem játszott. Mégis, a szokásos munkanap nyolc óra, és egy napi munka általában kisebb részekre bontható; az Agent-1-et úgy is tekinthetjük, mint egy szétszórtan dolgozót, aki gondos irányítás mellett jól boldogul.[29] Az okos emberek megtalálják a módját annak, hogyan automatizálják munkájuk rutinszerű részeit.[30]

Az OpenBrain vezetői az AI K+F automatizálásának egyik következményére is kitérnek: a biztonság fontosabbá vált. 2025 elején a legrosszabb forgatókönyv az algoritmikus titkok kiszivárgása volt; most, ha Kína ellopja az Agent-1 súlyait, közel 50%-kal növelhetik kutatási sebességüket.[31] Az OpenBrain biztonsági

---

28 A gyakorlatban arra számítunk, hogy az OpenBrain 8 hónapnál gyorsabb ütemben fogja kiadni a modelleket, de a rövidség kedvéért nem írjuk le az összes inkrementális kiadást.

29 Konkrétan, 80%-os pontszámot jósolunk az OSWorld-en (ami megfelel egy képzett, de nem szakértő embernek); 85%-ot a Cybench-en, ami megfelel a legjobb profi emberi csapatnak olyan hacker feladatokban, amelyekhez ezeknek a csapatoknak 4 órájuk van; és 1,3-at a RE-Bench-en, ami megfelel a legjobb szakértő embereknek, akiknek 8 órát adunk jól meghatározott AI-kutatási mérnöki feladatokra.

30 Az Agent-1 és utánzatai kereskedelmi szempontból sikeresek; 2025 folyamán az AI-vállalatok bevételei megháromszorozódnak, az OpenBrain értékelése pedig eléri az 1 000 000 dollárt. A Microsoft, a Google és az Amazon vezetésével az adatközpontokra fordított éves kiadások megduplázódnak 400 milliárd dollárra, és az USA több mint 5 GW AI teljesítményt vonz. További részletekért tekintse meg a Compute kiegészítő melléklet iparági mérőszámok című részét.

31 A mesterséges intelligencia biztonságával foglalkozó kutatók évek óta a legfontosabb veszélyes képességként tárgyalják a mesterséges intelligencia kutatás-fejlesztésének automatizálását. Elsődleges aggodalmuk az, hogy a belső telepítés az AI K+F felgyorsulásához vezet, ami megnehezíti az emberek számára, hogy lépést tartsanak a történésekkel és kitalálják, hogyan tegyék azokat biztonságossá. Az OpenBrain mint szervezet ezt pontosan az ellenkező irányba csavarja, és az AI K+F képesség veszélyes szintjére hivatkozva indokolja, hogy ne tájékoztassák a nyilvánosságot. Tehát a vállalat belső és nyilvános képességei közötti szakadék kezdetét látjuk. Az AI K+F az, amiben a mo-

szintje tipikusan egy gyorsan növekvő, ~3000 fős technológiai vállalatra jellemző, amely csak a képes kibercsoportok alacsony prioritású támadásai ellen védett (RAND SL2).[32] Keményen dolgoznak azon, hogy megvédjék súlyaikat és titkaikat a bennfentes fenyegetésektől és a vezető kiber bűnözői szindikátusoktól (SL3),[33] de a nemzetállamok elleni védelem (SL4&5) alig van a láthatáron[34].

2026 közepe: Kína felébred

Kínában a KKP kezdi megérezni az AGI.

A chipek exportjának ellenőrzése és a kormányzati támogatás hiánya miatt Kína a nyugati országokhoz képest alulfinanszírozott. A betiltott tajvani chipek csempészésével, régebbi chipek vásárlásával és az amerikai-tajvani határhoz képest körülbelül három évvel lemaradt hazai chipek gyártásával Kínának sikerült fenntartania a világ mesterséges intelligencia szempontjából releváns számítógépeinek körülbelül 12%-át - de a régebbi technológiával nehezebb dolgozni, és az ellátás állandó fejfájást okoz.[35] Néhány kiemelkedő alakulat, mint a DeepCent, nagyon lenyűgöző munkát végez korlátozott számítási kapacitással, de a számítási hiány korlátozza azt, amit kormányzati támogatás nélkül elérhetnek, és körülbelül hat hónappal vannak lemaradva a legjobb OpenBrain modellektől.[36]

A főtitkár régóta álmodott arról, hogy megduplázza a valós fizikai gyártást, és elkerülje az amerikai posztindusztriális dekadenciát. Gyanakvással tekintett a szoftvercégekre.[37] De a KKP sólymai arra figyelmeztetnek, hogy az AGI felé tartó növekvő versenyt már nem lehet figyelmen kívül hagyni. Így végül teljes mértékben elkötelezi magát a nagy AI-törekvés mellett, amelyet korábban megpróbált elkerülni. Elindítja a kínai mesterséges intelligencia kutatás államosítását, azonnali információ megosztási mechanizmust hozva létre a mesterséges intelligenciával foglalkozó vállalatok számára. Ez egy év alatt addig fokozódik, amíg a legjobb kutatók egy DeepCent által vezetett kollektívába tömörülnek, ahol megosztják egymással az algoritmikus meglátásokat, az adathalmazokat és a számítási erőforrásokat. A Tianwan erőműben (a világ legnagyobb atomerőmű-

---

dellek a legjobbak, ami ahhoz vezet, hogy a nyilvánosság egyre kevésbé érti az AI képességek határát.

32 Lásd: A Playbook for Securing AI Model Weights, RAND Corporation, 2024.

33 Az OpenBrain munkatársainak mintegy 5%-a a biztonsági csapatban dolgozik, és ők nagyon ügyesek, de a fenyegetések felülete is rendkívül nagy. Az sem segít, hogy ebben a szakaszban többnyire el vannak zárva attól, hogy olyan irányelveket hajtsanak végre, amelyek lassíthatják a kutatás előrehaladását. További részletekért lásd a Biztonsági előrejelzésünket.

34

35 Kínában jelenleg 3 millió H100e van, szemben az egy évvel ezelőtti 1,5 millióval 2025 közepén. További részletekért lásd a compute kiegészítő terjesztési szakaszát. Arra számítunk, hogy a csempészi erőfeszítések mintegy 60K GB300-at (450K H100e) biztosítanak, további 2M Huawei 910C-t gyártanak (800k H100e), és ~1M legálisan importált chipek (mint például az Nvidia H20 vagy B20) keveréke alkotja az utolsó 250K H100e-t.

36 Összehasonlításképpen: 2025 januárjában a DeepSeek kiadta az R1-et, amely versenyképes az OpenAI o1 modelljével, amelyet 2024 decemberében adtak ki. Úgy gondoljuk azonban, hogy a valódi különbség nagyobb, mint egy hónap, mivel az OpenAI valószínűleg már hónapok óta dolgozott az o1-en, és így feltehetően már néhány hónappal az o1 megjelenése előtt rendelkezett egy, az r1-hez nagyjából hasonló képességű előddel.

37 Ezzel szemben a CCP komolyan próbálta felgyorsítani a félvezető ipar.

vében) létrehoznak egy központosított fejlesztési zónát (CDZ), ahol a DeepCent új mega adatközpontja kap helyet, valamint rendkívül biztonságos lakó- és irodahelyiségek, ahová a kutatók végül átköltöznek. Kína mesterséges intelligencia szempontjából releváns számítógépeinek csaknem 50%-a már a DeepCent által vezetett kollektívának dolgozik,[38] és az új chipek több mint 80%-a a CDZ-be irányul.[39] A CDZ jelenleg rendelkezik a világ legnagyobb központosított klaszteréhez szükséges energiakapacitással.[40] Más párttagok szélsőséges intézkedéseket vitatnak meg a Nyugat chipelőnyének semlegesítésére. Tajvan blokádja? Teljes invázió?

De Kína a gyengébb modelljei miatt lemarad a mesterséges intelligencia algoritmusok terén. A kínai hírszerző ügynökségek - amelyek a világ legjobbjai közé tartoznak - megduplázzák az OpenBrain súlyainak ellopására irányuló terveiket. Ez egy sokkal összetettebb művelet, mint az algoritmikus titkok folyamatos, alacsony szintű orvvadászata; a súlyok egy több terabájtos fájl, amelyet egy rendkívül biztonságos szerveren tárolnak (az OpenBrain a RAND SL3-hez képestjavított biztonsággal rendelkezik). A kiberhadseregük úgy gondolja, hogy kémjeik segítségével sikerülhet nekik, de talán csak egyszer; az OpenBrain fel fogja fedezni a lopást, növelni fogja a biztonságot, és lehet, hogy nem kapnak több esélyt. Tehát (a CCP vezetése elgondolkodik) most kellene cselekedniük, és ellopni az Agent-1-et? Vagy várjanak egy fejlettebb modellre? Ha várnak, megkockáztatják, hogy az OpenBrain olyan mértékben növeli a biztonságot, hogy nem tudnak behatolni?

2026 vége: A mesterséges intelligencia megszüntet néhány állást

Éppen amikor úgy tűnt, hogy a többiek felzárkóznak, az OpenBrain ismét kiüti a versenytársakat az Agent-1-mini-modell kiadásával, amely 10x olcsóbb, mint az Agent-1, és könnyebben finomhangolható a különböző alkalmazásokhoz. A mesterséges intelligenciával kapcsolatos mainstream narratíva a „talán a hype el fog múlni" helyett „*gondolom, ez a következő nagy dolog*", de az emberek nem

[38] Eleinte a megosztás csak az ötletekre korlátozódik, és a vállalatok többnyire azzal vannak elfoglalva, hogy befejezzék az aktuális munkaterhelésüket, de az év végére a koordinációjuk annyira megnő, hogy már a modellsúlyokat is megosztják, sőt, a feladatokat is felosztják egymás adatközpontjai között. Például az egyik campuson egy új előképzési futtatást végeznek, egy másik a szintetikus adatok betáplálására, egy másik pedig mesterséges intelligencia-kutatási kísérletek futtatására összpontosít.

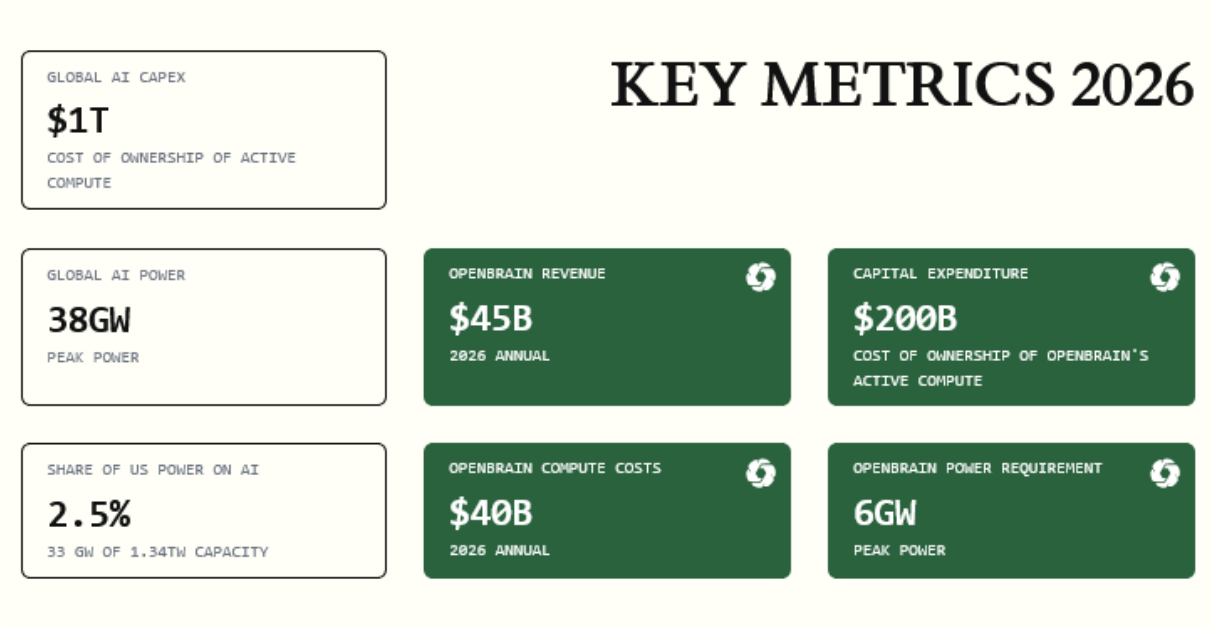
[39] Miért „csak" 50%-a Kína mesterséges intelligencia szempontjából releváns számítási kapacitásának? A többi számítási kapacitás a gazdaság egészét leköti, különösen az internetes és közösségi hálózati ajánló algoritmusokat. Ennek nagy része földrajzilag is túlságosan szétszóródott a sok kis klaszter között. Miért „csak" az új chipek 80%-a? Sok nagy megrendelést már más alkalmazásokhoz biztosítottak olyan vállalatok, amelyek nem akarnak 100%-ban hozzájárulni a nemzeti AI-erőfeszítésekhez. Az év végére azonban ez az arány eléri a 90%-ot.

[40] Még legalább egy év van hátra, hogy megkapják a chipeket, amelyekkel ezt a kapacitást ki tudják tölteni, és egy-két amerikai techóriás még mindig nagyobb decentralizált klaszterekkel fog rendelkezni.

értenek egyet abban, hogy mennyire nagy. Nagyobb, mint a közösségi média? Nagyobb, mint az okos telefonok? Nagyobb, mint a tűz?

A mesterséges intelligencia elkezdett munkahelyeket elvenni, de újakat is teremtett. A tőzsde 2026-ban 30%-ot emelkedett, az OpenBrain, az Nvidia és azon vállalatok vezetésével, amelyek a legsikeresebben integrálták az AI asszisztenseket. A junior szoftvermérnökök munkaerőpiaca felbolydult: az AI-k mindent meg tudnak csinálni, amit egy CS-diploma tanít, de azok az emberek, akik tudják, hogyan kell AI-csapatokat irányítani és minőség-ellenőrzést végezni, jól keresnek. Az üzleti guruk azt mondják az álláskeresőknek, hogy az AI ismerete a legfontosabb készség, amit az önéletrajzba kell írni. Sokan attól tartanak, hogy a mesterséges intelligenciák következő hulláma az ő munkahelyükért fog jönni; Washingtonban 10 000 ember tüntet az AI ellen.

A Védelmi Minisztérium (DOD) csendben elkezdi az OpenBrain közvetlen szerződtetését kiber-, adatelemzési és K+F feladatokra, de az integráció a bürokrácia és a DOD beszerzési folyamata miatt lassan halad.[41].



Miért nő jelentősen a bizonytalanságunk 2026 után

Előrejelzésünk a mai naptól 2026-ig lényegesen megalapozottabb, mint ami az alábbiakban következik. Ez részben azért van, mert ez közelebb van. De azért is, mert a mesterséges intelligencia világra gyakorolt hatásai 2027-ben kezdenek igazán felerősödni. A 2025-ös és 2026-os előrejelzésünk nagymértékben a számítási kapacitásnövelés, az algoritmikus fejlesztések és a benchmarkok teljesítményének extrapolálásából származik. A forgatókönyvnek ezen a pontján kezdjük látni a mesterséges intelligen-

[41] Ez egy egyéb tranzakciós hatósági (OTA) szerződésen keresztül történik, és magas prioritású DX minősítést kap. Egyáltalán nem vagyunk biztosak abban, hogy ez a legvalószínűbb módja az együttműködés finanszírozásának, de a konkrétság kedvéért választottunk valami konkrétat. A szerződést nyilvánosan bejelentették, de az OpenBrain kommunikációjában nem hangsúlyozzák.

cia által felgyorsított mesterséges intelligencia-kutatás és -fejlesztés jelentős hatásait az idővonalon, ami arra késztet minket, hogy felfelé módosítsuk a trendvonalakra vonatkozó becsléseinket. De ezek a dinamikák természetüknél fogva sokkal kevésbé kiszámíthatóak.

A 2027-es év folyamán az AI-k az OpenBrain kutatómérnökeinek munkáját többnyire elvégzőktől az összes feladatban az emberek háttérbe szorításáig fejlődnek. Ez nagyjából a mi medián becslésünket jelenti, de úgy gondoljuk, hogy hihető, hogy ez akár ~5x lassabban vagy gyorsabban történik.

Háttérkutatásainkért lásd az idővonalak előrejelzését és a felszállási előrejelzést. Ezeket alább bővíthető táblázatokban is összefoglaljuk.

2027 január: A 2. ügynök sosem fejezi be a tanulást

Az OpenBrain az Agent-1 segítségével most az Agent-2-t képzi tovább. Minden eddiginél nagyobb hangsúlyt fektet a kiváló minőségű adatokra. Rengeteg szintetikus adatot állítanak elő, értékelnek és szűrnek a minőség szempontjából, mielőtt az Agent-2-t betáplálják.[42] Ezen felül több milliárd dollárt fizetnek emberi munkaerőnek, hogy rögzítsék, amint hosszú távú feladatokat oldanak meg.[43] *Mindennek* tetejébe szinte folyamatosan képzik az Agent-2-t megerősített tanulással, változatos, nehéz feladatok egyre bővülő halmazán: sok videojátékon, sok kódolási kihíváson, sok kutatási feladaton. Az Agent-2, még inkább, mint a korábbi modellek, gyakorlatilag „online tanulás", mivel úgy építették fel, hogy soha nem fejezi be a képzést. A súlyok minden nap frissülnek a legújabb verzióra, amelyet az előző nap az előző verzió által generált több adaton képeztek ki.

Az Agent-1-et az AI K+F feladatokra optimalizálták, remélve, hogy intelligencia-robbanást indít el.[44] Az OpenBrain megduplázza ezt a stratégiát az Agent-2-vel. Minőségileg majdnem olyan jó, mint a legjobb emberi szakértők a kutatás-fejlesztésben (kísérletek tervezése és végrehajtása), és olyan jó, mint az OpenBrain 25. százalékos tudósa a „kutatási ízlésben" (annak eldöntése, hogy mit tanulmányozzanak legközelebb, milyen kísérleteket futtassanak, vagy hogy milyen lehetséges új paradigmákról van sejtésük).[45] Míg a legutóbbi Agent-1 meg tudta duplázni az OpenBrain algoritmikus fejlődésének ütemét, az Agent-2 már képes megháromszorozni azt, és idővel tovább fog fejlődni. A gyakorlatban ez úgy néz ki, hogy minden OpenBrain-kutató egy AI-„csapat" „menedzserévé" válik.

---

42 Például visszautasított mintavételezéssel, ahogyan azt a DeepSeek R1 és a Meta Llama 3.1 tette.

43 Nagyjából 20 000 teljes munkaidős egyenérték * 100 $/óra * 2000 óra/év = 4milliárd $/év.

44 Vagyis azt remélik, hogy az Agent-2 munkaerő felhasználásával jelentősen fel tudják gyorsítani a saját alapvető kutatási tevékenységüket, így gyorsabban tudnak jobb AI-kat képezni, amelyek további gyorsulást okozhatnak, stb. Ennek szélsőséges (és szerintünk mégis meglepően hihető) változatának elemzését lásd ebben a jelentésben.

45 Van némi bizonyíték arra, hogy az AI-rendszerek már rendelkeznek némi alapkutatási ízléssel, pl. ez a cikk egy AI-társ-tudósról és ez a cikk az LLM-fókuszú kutatási dolgozatok újszerű ötleteinek generálásáról. A jobb ízléssel rendelkező kutatócsoportoknak még mindig kísérleteket kell végezniük, de jobban megválasztják, hogy milyen kísérleteket futtassanak, és milyen tanulságokat vonjanak le az eredményekből.

Az új képességekkel új veszélyek is járnak. A biztonsági csapat megállapítja, hogy ha az Agent-2 valahogy megszökne a cégtől, és autonóm módon „túlélni" és „replikálódni" akarna, akkor képes lenne erre. Vagyis autonóm módon kidolgozhatna és végrehajthatna olyan terveket, amelyekkel feltörhetné az AI szervereit, másolatokat telepíthetne magáról, elkerülhetné a felderítést, és ezt a biztonságos bázist felhasználhatná bármilyen más céljának elérésére (bár az, hogy ezt a hetek múlásával mennyire hatékonyan tenné, ismeretlen és kétséges). Ezek az eredmények csak azt mutatják, hogy a modell képes elvégezni ezeket a feladatokat, nem pedig azt, hogy „akarja-e" ezt tenni. Mégis, már az is nyugtalanító, hogy tudjuk, ez lehetséges.

Tekintettel az új modell „veszélyeire", az OpenBrain „felelősségteljesen" úgy dönt, hogy egyelőre nem hozza nyilvánosságra (valójában a belső AI K+F-re akarnak koncentrálni).[46] Az Agent-2 teljes képességeinek ismerete egy elit silóra korlátozódik, amely a közvetlen csapatot, az OpenBrain vezetőségét és biztonságát, néhány tucat amerikai kormányzati tisztviselőt és az OpenBrainbe évek óta beszivárgó CCP kémek légióit tartalmazza.[47]

2027 februárja: Kína ellopja az Agent-2-t

Az OpenBrain bemutatja az Agent-2-t a kormánynak, beleértve a Nemzetbiztonsági Tanácsot (NSC), a Védelmi Minisztériumot (DOD) és az Egyesült Államok Mesterséges Intelligencia Biztonsági Intézetét (AISI). Az OpenBrain jó kapcsolatot akar fenntartani a végrehajtó hatalommal, mert gyakorlatilag ez az egyetlen szereplő, aki most megállíthatja őket, és ha nem tájékoztatják, akkor az úgyis rájöhet, és elégedetlen lesz.[48]

A tisztviselőket leginkább a kiber hadviselési képességei érdeklik: Az Agent-2 „*csak*" egy kicsit rosszabb, mint a legjobb emberi hackerek, de több ezer példánya párhuzamosan futtatható, gyorsabban keresve és kihasználva a gyenge pontokat, mint ahogy a védők reagálni tudnak. A Védelmi Minisztérium ezt kritikus előnynek tartja a kiber hadviselésben, és a mesterséges intelligencia a kormány prioritási listájának 5. helyéről a 2. helyre kerül.[49] Valaki megemlíti az OpenBrain államosításának lehetőségét, de más kabineti tisztviselők szerint ez

---

46 Továbbá, még mindig nagy visszatérést látnak az edzés utáni folytatásra, és így nincs különösebb okuk arra, hogy most kiadjanak egy terméket. Ráadásul a belső eredmények és a modell kiadása között gyakran néhány hónapos késedelem van (ennek egy részét a red teamingre fordítják).

47 A silóban 200 OpenBrain kutató (10 vezető, 140 a képességekkel foglalkozó csapatból, 25 a biztonsági, felügyeleti és ellenőrzési csoportból, 15 a felkészültségi- vagy RSP- típusú csapatokból és 10 az összehangolásból) és 50 kormányzati tisztviselő (15 a Fehér Házból, 5 az AISI-tól, 10 a Védelmi Minisztériumból, 10 a DOE-től és 10 a CISA-tól) dolgozik. A projektben több kém is részt vesz, különböző mértékű hajlandósággal. Ők gyakran kényszerített személyek (pl. Kínában élő családdal), nem pedig önkéntes munkatársak.

48 Miért nem tudja a Kongresszus megállítani őket? Mi a helyzet a bíróságokkal? Túl lassúak, túlságosan megosztottak és/vagy a POTUS-nak engedelmeskednek, túlságosan kívül esnek a körön. Emellett pozitív okai is vannak annak, hogy jóban legyünk a Fehér Házzal: segíthet a bürokrácia csökkentésével, a biztonsággal kapcsolatos segítséggel, az állami törvényhozás megelőzésével, és talán még a versenytársaik lelassítását célzó intézkedésekkel is.

49 #Az 1. számú ország éppen az aktuális válsággal foglalkozik - talán egy Tajvan körüli fegyverkezés, egy új háború a Közel-Keleten vagy valamilyen belpolitikai felfordulás.

korai lenne. Egy munkatárs feljegyzést készít, amelyben bemutatja az elnöknek a lehetőségeket, a szokásos módon történő működésétől a teljes államosításig. Az elnök a tanácsadóira, a technológiai ipar vezetőire hagyatkozik, akik szerint az államosítás „*megölné az aranytojást tojó tyúkot*". Úgy dönt, hogy egyelőre nem tesz nagyobb lépéseket, és csak további biztonsági követelményeket ír elő az OpenBrain és a Védelmi Minisztérium közötti szerződéshez.

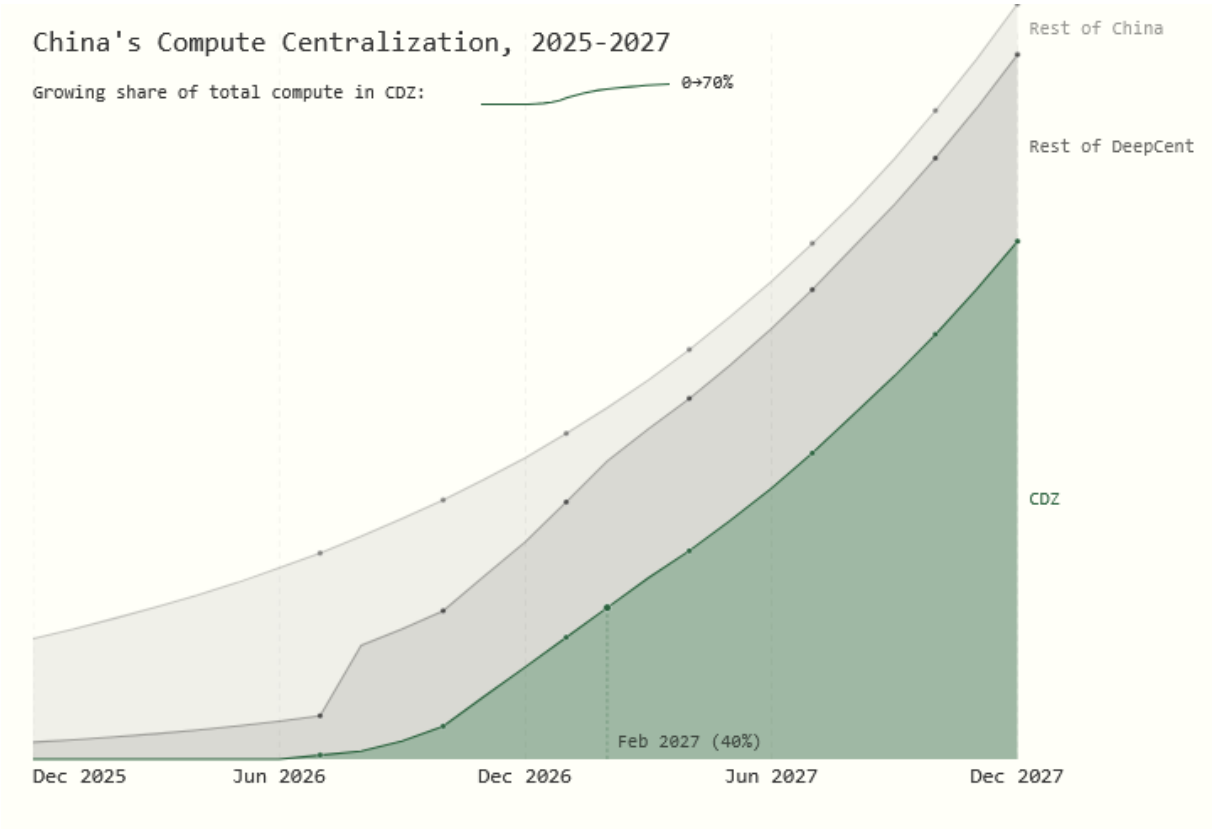
A változtatások túl későn jönnek. A CCP vezetése felismeri az Agent-2 fontosságát, és utasítja kémjeit és kiber hadseregét, hogy lopják el a súlyokat. Egy kora reggel az Agent-1 forgalomfigyelő ügynök rendellenes átvitelt észlel. Riasztja a vállalat vezetőit, akik értesítik a Fehér Házat. A nemzetállami szintű művelet jelei félreérthetetlenek, és a lopás felerősíti a fegyverkezési verseny érzését.

Az Agent-2 modell súlyainak ellopása

Úgy gondoljuk, hogy a kínai hírszerzés már évek óta különböző módon kompromittálta az OpenBrain-t, és valószínűleg naprakészen tartotta az algoritmikus titkokat, sőt időről időre kódot is lopott, mivel azt sokkal könnyebb megszerezni, mint a súlyokat, és sokkal nehezebb felderíteni.
A súlyok ellopását úgy képzeljük el, mint egy sor összehangolt, kisebb (azaz gyors, de nem titkos) lopást egy sor Nvidia NVL72 GB300 szerveren, amelyeken az Agent-2 súlyok másolatai futnak. A szervereket a törvényes alkalmazotti hozzáféréssel veszélyeztetik (egy barátságos, kényszerített vagy akaratlan bennfentes adminisztrátori hitelesítő adatokkal segíti a CCP lopási erőfeszítéseit). Annak ellenére, hogy az Nvidia bizalmas számítási rendszerének megerősített verziójával fut, a bennfentes hitelesítő adatok admin szintű jogosultságokat biztosítanak a támadónak (beleértve a bizalmas VM irányítását a biztonságos enklávén belül), lehetővé téve számukra, hogy többszörös koordinált súlyok átvitelét kezdeményezzék kis 4%-os töredékekben (100 GB-os darabokban) 25 különböző szerverről.
Az Nvidia protokolljaiban a memóriában (HBM) lévő plaintext súlyok titkosítva vannak, mielőtt továbbítják őket, de a támadók éppen azon a szerveren belül vannak, amely ismeri ezt a privát (szimmetrikus Diffie-Hellman) kulcsot, így nem kell aggódniuk a helyszíni dekódolás miatt (ami valószínűleg riadót keltene), és csak a titkosított súlyokat a szerver front-end hálózati kártyáin keresztül kiszivárogtatják. A teljes adatközpont kilépési sávszélessége (az adatok távozási sebessége) 100 GB/másodperc tartományban van, így a ~2,5 TB-os súlyfájl (5T félpontosságú paraméterek) kiszivárgásának szerverenként 1 GB/s alatti értékre történő korlátozásával elkerülhető a hálózati forgalom jelentős megugrása, és lehetővé teszi, hogy az egyes fájlok egyenként kevesebb mint 5 perc alatt elhagyják az adatközpontot az üvegszálas kábeleken keresztül. Az élő megfigyelést vagy megtévesztik a támadóknak a fájlok elrejtésére és felosztására irányuló erőfeszítései, vagy pedig egyenesen letiltják.
A súlyokat ezután különböző párhuzamos csatornákon és IP-maszkoló rétegeken keresztül Kínába irányítják, hogy ott helyben dekódolják őket az

ellopott titkos kulcs(ok)kal. A művelet teljes aktív része (az első szerver kompromittálásától a súlyok teljes kiszivárgásáig) kevesebb mint két óra alatt befejeződik.
Oroszország ezen a ponton szintén megpróbálja ellopni a modellt, de kudarcot vall - túl sokáig vártak, és nem a megfelelő helyekre való beszivárgásra fordították az erőforrásokat. A kémeik rutinszerűen lopnak algoritmikus titkokat az amerikai AI cégektől, de saját, figyelemre méltó AGI projekt nélkül ezek nem túl hasznosak.

A Fehér Ház rövidebb pórázra veszi az OpenBrain-t, és katonai és hírszerzési személyzettel bővíti biztonsági csapatát. Elsődleges céljuk a további súlylopások megakadályozása.[50]



A lopás megtorlásaként az elnök kiber támadásokat engedélyez a DeepCent szabotálására. Mostanra azonban Kína a mesterséges intelligencia szempontjából releváns számítógépeinek[51] 40%-a a CDZ-ben van, ahol agresszívan megkeményí-

50 A legegyszerűbb megoldás az lenne, ha a vállalati adatközpontokból minden nagy sávszélességű kapcsolatot lezárnának, de ez a nagyméretű fájlátvitelt a gyakorlatban nem megvalósítható mértékben lelassítaná. Ehelyett a legtöbb külső kapcsolatot le tudják zárni, de a képzésben aktívan részt vevő adatközpontoknak gyorsan kell súlyokat cserélniük egymással; e kapcsolatok fojtása túlságosan hátráltatná a fejlődést. Az OpenBrain tehát fokozott felügyelet és egy extra titkosítási réteg segítségével tartja fenn ezeket a kapcsolatokat.

51 Emlékezzünk arra, hogy 2026 közepe óta Kína az újonnan beszerzett mesterséges intelligenciachipek 80%-át a CDZ-be irányította. Tekintettel arra, hogy a számítási kapacitásuk 2026 eleje óta megduplázódott (a globális termelési trenddel összhangban), ez a CDZ-t 2M 2024-egyenértékű GPU-ra (H100) és 2 GW energiafelvételre teszi. Az OpenBrain még mindig a DeepCent számítási ka-

tették a biztonságot az airgapping (külső kapcsolatok lezárása) és a belső silózás révén. A műveletek nem okoznak komoly, azonnali károkat. A feszültség fokozódik, mindkét fél jelzi komolyságát azzal, hogy katonai eszközöket helyez át Tajvan körül, a DeepCent pedig igyekszik az Agent-2 hatékony működésbe hozni, hogy megkezdhesse a mesterséges intelligencia kutatásának fellendítését[52].

2027. március: Algoritmikus áttörések

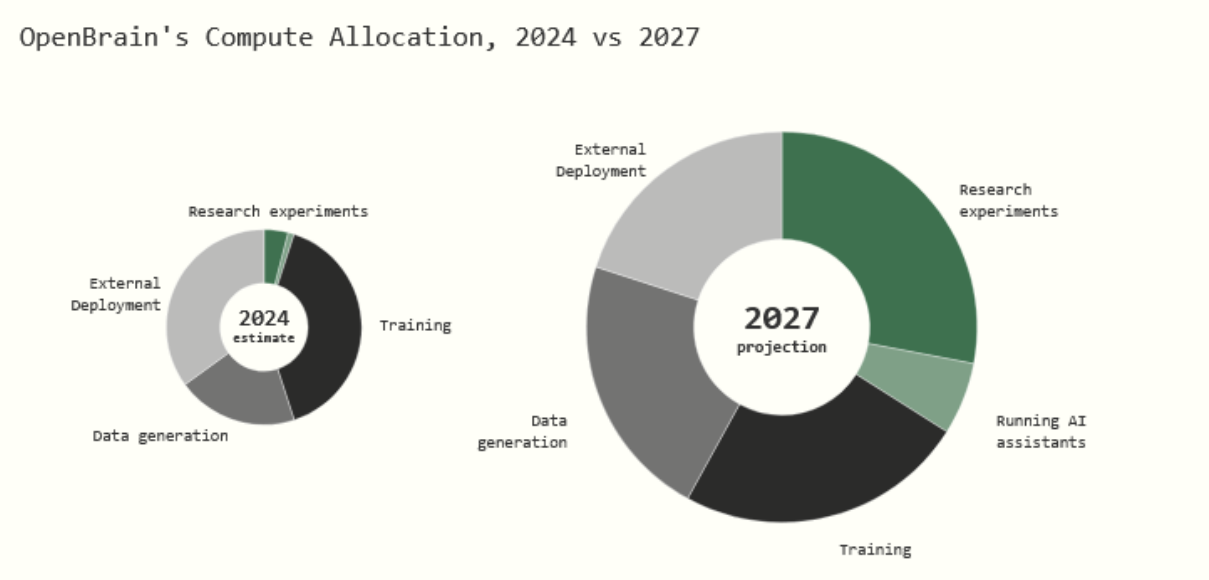
Három hatalmas, Agent-2 másolatokkal teli adatközpont dolgozik éjjel-nappal, szintetikus képzési adatokat termelve. További kettő a súlyok frissítésére szolgál. Az Agent-2 napról napra okosabbá válik.

Az OpenBrain az Agent-2 több ezer automatizált kutatójának segítségével jelentős algoritmikus előrelépéseket tesz. Az egyik ilyen áttörés a mesterséges intelligencia szövegalapú scratchpadjának (gondolatlánc) kiegészítése egy nagyobb sávszélességű gondolkodási folyamattal (neurális ismétlődés és memória). Egy másik a nagy erőfeszítéssel járó feladatmegoldások eredményeiből való tanulás skálázhatóbb és hatékonyabb módja (iterált desztilláció és erősítés).

Az ezeket az áttöréseket magában foglaló új mesterséges intelligencia rendszer az Agent-3 nevet kapta.

pacitásának kétszeresével rendelkezik, és a többi amerikai vállalat együttesen 5x annyival, mint ők. További részletekért lásd a compute-kiegészítés elosztási szakaszát.

52 A folyamatban lévő nemzeti központosítás ellenére a DeepCentnek még mindig van egy marginális, de fontos számítási hátránya. Amellett, hogy Kínának körülbelül feleannyi teljes feldolgozási teljesítménye van, több teljes chipet kell használnia, amelyek (átlagosan) rosszabb minőségűek, és heterogén GPU-kat (amelyeket nem mindig könnyű hatékonyan összekapcsolni), amelyek mindkettő megterheli a chipek közötti hálózatépítést. Vannak szoftveres különbségek is (pl. a nem Nvidia-GPU-k nem rendelkeznek CUDA-val), valamint a hardveres specifikációkban mutatkozó különbségek, ami azt jelenti, hogy a képzési kódjuk bonyolultabb, lassabb és hibalehetőségekkel jár. A magas kihasználtság elérése egy downstream kihívás, az adatbevitel, az ütemezés, a kollektív kommunikáció és a párhuzamossági algoritmusok elmaradnak az amerikai vállalatoktól. Ezeknek a problémáknak az enyhítése azonban leginkább erőfeszítés és tesztelés kérdése, ami nagyszerű feladatot jelent az újonnan ellopott Agent-2 számára, és körülbelül egy hónapon belül a kínai projekt üzemideje és átlagos erőforrás-kihasználtságuk a képzési és következtetési munkaterhelések között javul, és már csak minimálisan marad el az amerikaiaktól.



Neurális ismétlődés és memória

A neurális rekurzió és memória lehetővé teszi, hogy a mesterséges intelligencia modellek hosszabb ideig gondolkodjanak anélkül, hogy ezeket a gondolatokat szövegként kellene leírni.
Képzeld el, hogy egy ember rövid távú memóriavesztéssel küzd, úgy, hogy folyamatosan papírra kell írnod a gondolataidat, hogy néhány perc múlva tudd, mi történik. Lassan és fájdalmasan haladhatnál a matematikai feladatok megoldásában, a kódírásban stb. de sokkal könnyebb lenne, ha közvetlenül emlékezhetnél a gondolataidra anélkül, hogy le kellene írnod, majd el kellene olvasnod őket. Ezt hozza a neurális rekurzió és a memória a mesterséges intelligencia modellekbe.
Szakszerűbben fogalmazva:
A hagyományos figyelemmechanizmusok lehetővé teszik, hogy egy modellben a későbbi előrehaladások a modell korábbi tokenekre vonatkozó közbenső aktivációit lássák. *Visszafelé* (a későbbi rétegekből a korábbi rétegekbe) azonban csak a tokeneken keresztül tudnak információt átadni. Ez azt jelenti, hogy ha egy hagyományos nagy nyelvi modell (LLM, pl. a GPT modellsorozat) olyan következtetési láncot akar végrehajtani, amely több soros műveletet igényel, mint ahány réteg van a modellben, a modell kénytelen információt tokenekbe foglalni, amelyeket aztán vissza tud továbbítani önmagába. Ez azonban rendkívül korlátozó - a tokenek csak nagyon kis mennyiségű információt tudnak tárolni. Tegyük fel, hogy egy LLM szókincsének mérete ~100 000, akkor minden token $\log_2(100k)=16,6$ bit információt tartalmaz, ami körülbelül egy lebegőpontos szám mérete (feltételezve, hogy FP16-ban képezzük). Eközben a reziduális adatfolyamok - amelyeket az LLM-ben a rétegek közötti információ átadására használnak - több ezer lebegőpontos számot tartalmaznak.

Ezt a szűk keresztmetszetet elkerülhetjük **neurális** módszerrel : az LLM reziduális folyamát (amely több ezer dimenziós vektorból áll) visszaadjuk a modell korai rétegeinek, így egy nagydimenziós gondolatmenetet kapunk, amely potenciálisan több mint 1000-szer több információt továbbít.

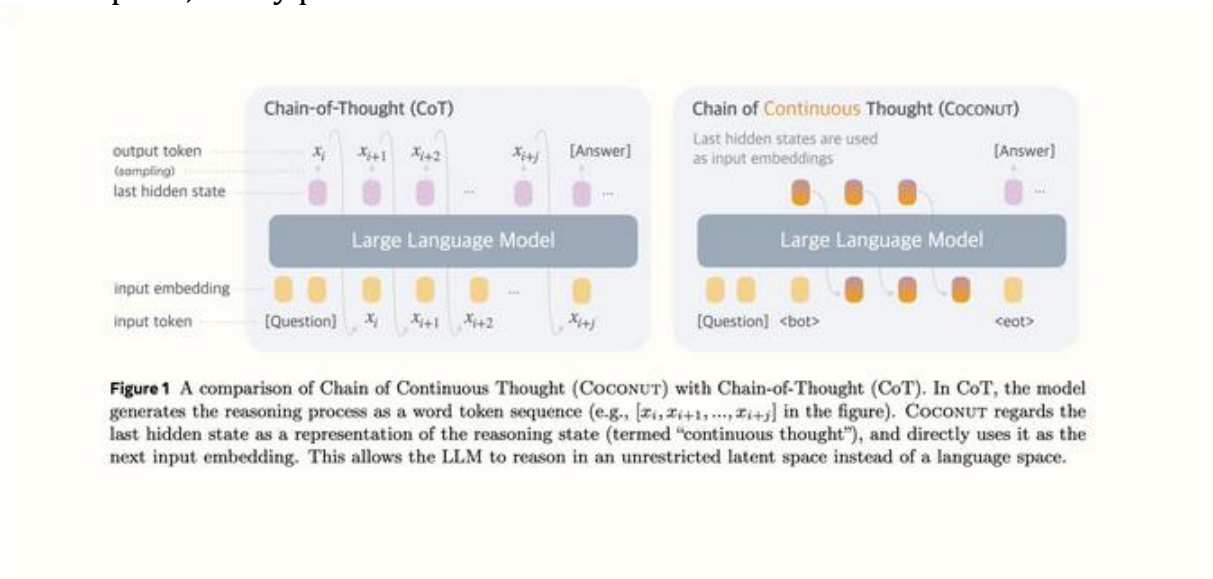


Figure 1 A comparison of Chain of Continuous Thought (Coconut) with Chain-of-Thought (CoT). In CoT, the model generates the reasoning process as a word token sequence (e.g., $[x_i, x_{i+1}, ..., x_{i+j}]$ in the figure). Coconut regards the last hidden state as a representation of the reasoning state (termed "continuous thought"), and directly uses it as the next input embedding. This allows the LLM to reason in an unrestricted latent space instead of a language space.

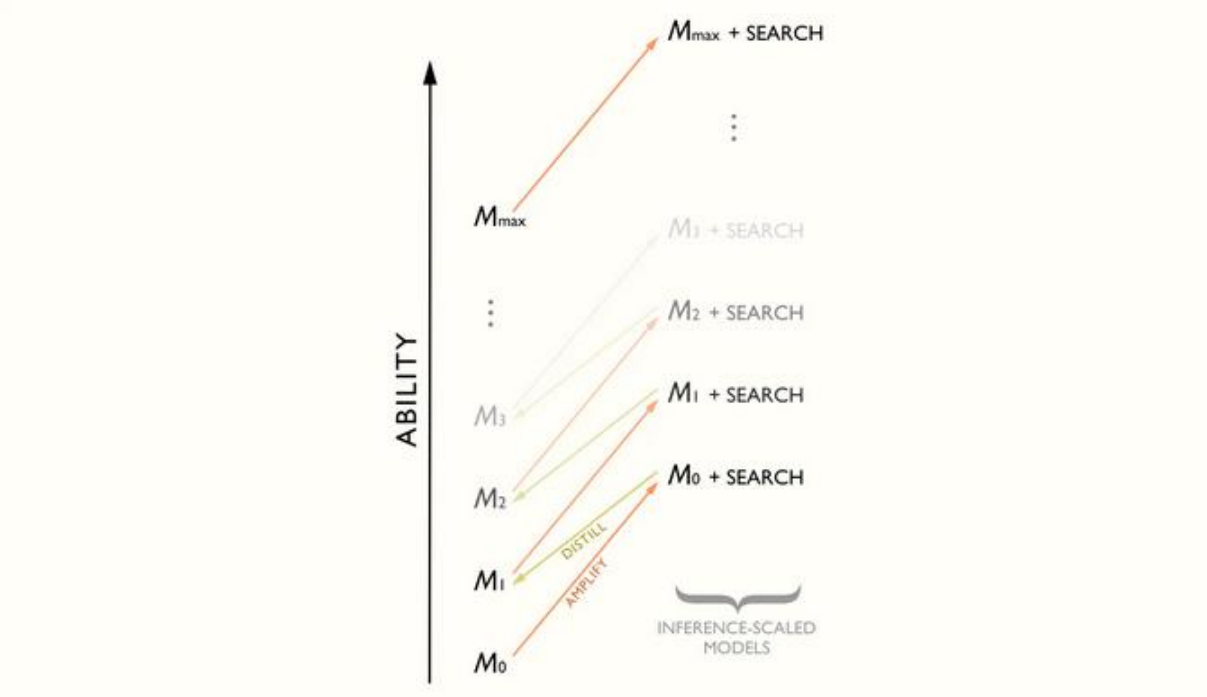
Az ábra Hao et al. 2024-es, ezt az ötletet megvalósító Meta-dolgozatából származik.
Azért nevezzük ezt „neurális nyelvnek", mert az angol szavakkal ellentétben ezeket a magas dimenziós vektorokat az ember valószínűleg meglehetősen nehezen tudja értelmezni. A múltban a kutatók jó ötletet kaphattak arról, hogy mit gondolnak az LLM-ek egyszerűen a gondolatláncának elolvasásával. Most a kutatóknak meg kell kérniük a modellt, hogy fordítsa le és foglalja össze gondolatait, vagy a neuralese-t a korlátozott értelmezhetőségi eszközeikkel kirakják.
Hasonlóképpen, a régebbi AI chatbotok és ügynökök külső szövegalapú memóriabankokkal rendelkeztek, mint egy ember, aki papíron jegyzetel. Az új mesterséges intelligencia hosszú távú memóriája szöveg helyett vektorok kötege, így gondolatai tömörebbek és magasabb dimenziójúak. A memóriabankoknak több típusa létezik; egyeseket ideiglenesen használnak egyetlen, több lépést tartalmazó feladathoz, másokat egy személy, vállalat vagy munkatípus (pl. programozás) által használt összes ágens megosztva használ.
Tudomásunk szerint a vezető AI-cégek, mint a Meta, a Google DeepMind, az OpenAI és az Anthropic még nem valósították meg ténylegesen ezt az ötletet a határmodelljeikben. Feltételezésünk szerint azért, mert a teljesítménynövekedés kicsi a bevezetett képzési ineffektivitásokhoz képest. A modell edzés előtti fázisában és az edzés utáni felügyelt tanulás során, például az utasítások finomhangolásakor, hogy bizonyos válaszokat produkáljanak , a hatékonysági hiányosságok abból erednek, hogy nem tudnak sok tokent párhuzamosan megjósolni, ami rosszabb GPU-kihasználtsághoz vezet. Neurális elemzés nélkül a modell egyszerre meg tudja jósolni az „Ez egy példa" mondat egészét, mivel már tudja, hogy az

„is” generálásához a bemenet „This” lesz, az „an” bemenet „This is” lesz, stb. A neurálisnál azonban nem tudni, hogy a „This” generálása után mi lesz a neurális vektor, amin keresztül a következő tokenhez jutunk. Ezért minden egyes tokent egyesével kell megjósolni. Az, hogy nem lehet az összes tokent párhuzamosan megjósolni, csökkenti a hatékonyságot azokban az esetekben, amikor az összes token előre ismert. Arra a kérdésre, hogy miért nem került még neuralese az utótanításba, azt feltételezzük, hogy a jelenlegi technikákkal a nyereség részben azért korlátozott, mert az utótanítás a folyamat kis részét képezi. Előrejelzésünk szerint 2027 áprilisára a költség-haszon arány sokkal jobbnak tűnik a neuralese esetében, mivel jobb technikákat fejlesztettek ki, és a képzés nagyobb hányada a képzés utáni képzés.
Ha ez nem következik be, akkor is történhetnek más dolgok, amelyek a történetünk szempontjából funkcionálisan hasonlóan végződnek. Például lehet, hogy a modelleket arra képzik ki, hogy olyan mesterséges nyelveken gondolkodjanak, amelyek hatékonyabbak a természetes nyelvnél, de az ember számára nehezen értelmezhetőek. Vagy talán bevett gyakorlattá válik, hogy az angol gondolatláncokat úgy képezzék ki, hogy szépen nézzenek ki, így a mesterséges intelligenciák ügyesek lesznek abban, hogy finoman kommunikáljanak egymással olyan üzenetekben, amelyek a monitorok számára jóindulatúnak tűnnek.
Ugyanakkor az is lehetséges, hogy a mesterséges intelligenciák, amelyek elsőként automatizálják a mesterséges intelligencia kutatását és fejlesztését, még mindig többnyire hű angol gondolatmenetekben gondolkodnak majd. Ha ez így van, akkor sokkal könnyebb lesz észrevenni a félreértéseket, és összességében a történetünk lényegesen más és optimistább lenne.

Iterált desztilláció és erősítés (IDA)

Az általános intelligencia önfejlesztése korábban kisebb sikereket ért el. De 2027 elején hatalmas hozamokat lát. Az IDA, az ehhez szükséges két összetevő:

1. **Az erősítés**: Adott egy M0 modell, költsünk több erőforrást a teljesítmény javítására. Például azzal, hogy a modell hosszabb ideig gondolkodhat, vagy sok példányt futtathat párhuzamosan, vagy mindkettőt, valamint azzal, hogy az eredmény kiértékelésére és a legjobb válaszok kiválasztására is hasonlóan intenzív folyamatot alkalmazunk, nagyságrendekkel több számítási időt fordíthatunk arra, hogy észrevehetően jobb minőségű válaszokat (vagy munkatermékeket) kapjunk. Nevezzük ezt a drága rendszert Amp(M0)-nak.
2. **Lepárlás**: Adott egy felerősített modell Amp(M0), képezzünk ki egy új modellt M1-et, hogy utánozza azt, azaz hogy ugyanolyan eredményt kapjunk, mint Amp(M0), de gyorsabban és kevesebb számítással. Az eredmény remélhetőleg egy okosabb modell, az M1 lesz. Ezután megismételheti a folyamatot.



Az IDA vizualizációja Ordtól, 2025.

Az AlphaGo-t így képezték ki: Monte-Carlo fakereséssel és önjátszással, mint *erősítő* lépéssel, és megerősítő tanulással, mint *desztilláló* lépéssel. Ez emberfeletti teljesítményt eredményezett a Go-ban. Most azonban az Agent-3 képes ezt kihasználni, hogy emberfeletti teljesítményt érjen el a kódolásban.

1. Az erősítési lépés úgy működik, hogy az Agent-3 hosszabb ideig gondolkodik, eszközöket használ, vagy konzultál más mesterséges intelligenciákkal. Ilyenkor gyakran rájön, hogy hibát követett el, vagy új felismeréssel áll elő. Ez nagy mennyiségű képzési adatot eredményez: a kutatási kísérletek címkézett pályáját azzal, hogy sikerrel jártak-e vagy sem. Ide tartoznak az olyan technikák is, mint az ellenőrizhető feladatokon a Best of N, majd a legjobb pályák megtartása.
2. A desztillációs lépés policy-gradiens megerősítő tanulási algoritmusokat használ, hogy a modell internalizálja a felerősített érvelést. Ezen a ponton az OpenBrain jobb RL algoritmusokat fedezett fel a proximal policy optimization (PPO) érájában. Folyamatosan desztillálják azt, amire az Agent-3 sok gondolkodás után egyetlen lépésben képes következtetni, ami folyamatosan javítja azt, amit egyetlen lépésben képes gondolni, és így tovább.

Az IDA korai változatai sok éven át dolgoztak könnyen ellenőrizhető feladatokon, például olyan matematikai és kódolási problémákon, amelyekre egyértelmű válasz van, mivel a modellek felerősítésére használt technikák gyakran támaszkodnak a pontosság valamilyen alapigazságjeléhez való hozzáférésre.

Mostanra a modellek elég jóvá váltak szubjektívebb dolgok (pl. egy munkatermék minősége) ellenőrzésére, ami lehetővé teszi az IDA használatát a modell javítására számos feladatnál.

Az új képességek áttörésének köszönhetően az Agent-3 gyors és olcsó emberfeletti kódoló. Az OpenBrain 200 000 Agent-3 másolatot futtat párhuzamosan, ami a legjobb emberi kódoló 50 000 másolatának megfelelő munkaerőt jelent, 30-szorosára gyorsítva.[53] Az OpenBrain továbbra is megtartja emberi mérnökeit, mert ők rendelkeznek az Agent-3 másolatok csapatainak irányításához szükséges kiegészítő készségekkel. Például a kutatási ízlés a hosszabb visszacsatolási hurkok és a kevesebb adat elérhetősége miatt nehezen képezhetőnek bizonyult.[54] Ez a hatalmas emberfeletti munkaerő „csak" 4x gyorsítja fel az OpenBrain algoritmikus fejlődésének általános ütemét a szűk keresztmetszetek és a kódolói munka csökkenő hozama miatt.[55]

Most, hogy a kódolás teljesen automatizált, az OpenBrain gyorsan képes kiváló minőségű képzési környezeteket gyártani, hogy megtanítsa az Agent-3 gyenge képességeit, mint például a kutatási ízlést és a nagyszabású koordinációt. Míg a korábbi képzési környezetek a „Itt van néhány GPU és a kísérletekre vonatkozó utasítások a kódoláshoz és futtatáshoz, a teljesítményedet úgy értékeljük, mintha ML mérnök lennél", most a „*Itt van néhány száz GPU, egy internetkapcsolat és néhány kutatási kihívás; neked és ezer másik példánynak együtt kell dolgoznod a kutatási eredmények érdekében. Minél lenyűgözőbb, annál magasabb a pontszámod*".

Miért jósolunk emberfeletti kódolót 2027 elejére

Időrendi előrejelzésünkben, megjósoljuk, hogy az OpenBrain mikor fog belsőleg kifejleszteni egy *emberfeletti kódolót (SC)*: egy olyan mesterséges intelligencia rendszert, amely képes bármilyen kódolási feladatot elvégezni, amit a legjobb AGI cég mérnöke, miközben sokkal gyorsabb és olcsóbb.

A METR egy nemrégiben készült jelentése szerint a mesterséges intelligenciák által megoldható kódolási feladatok hossza, az „*időhorizontjuk*", 2019 és 2024 között 7 havonta, 2024-től pedig 4 havonta megduplázódik. Ha a tendencia tovább gyorsul, 2027 márciusára a mesterséges intelligenciák 80%-os megbízhatósággal oldhatnak meg olyan szoftverfeladatokat, amelyek elvégzése egy képzett embernek éveket venne igénybe.

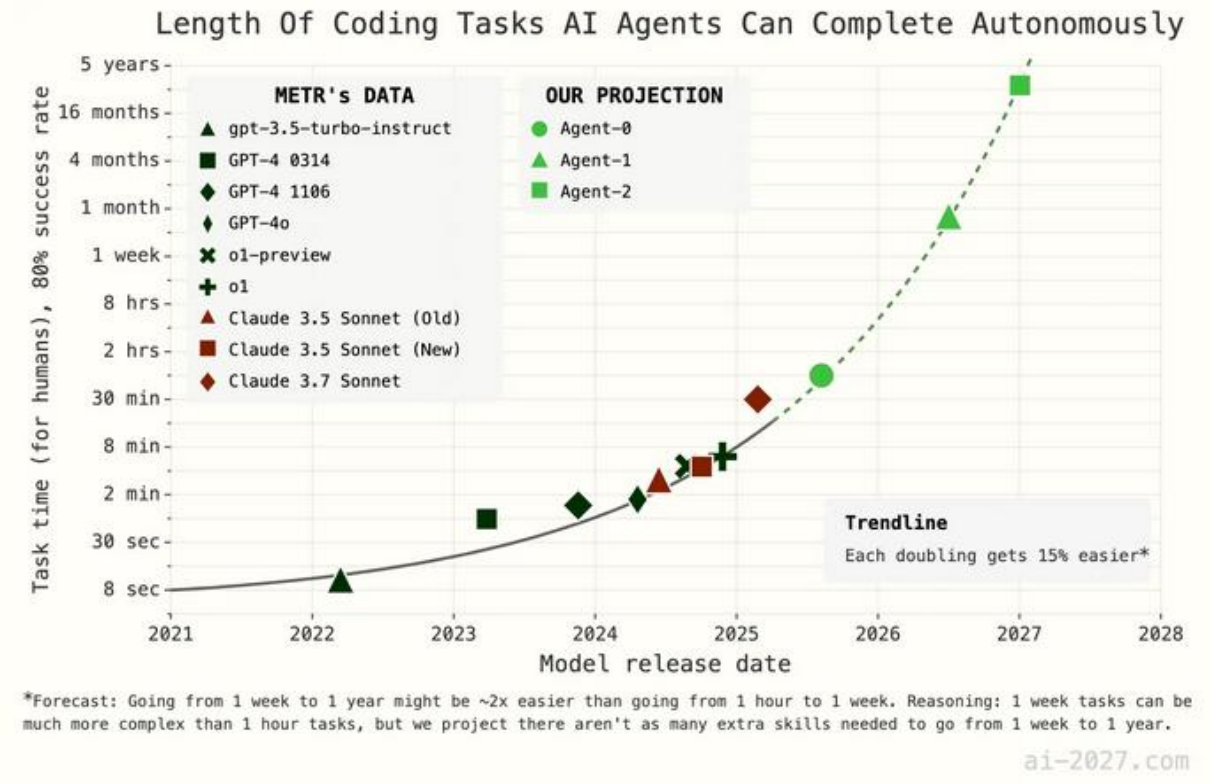
---

53 Arra számítunk, hogy az Agent-3 ma egy nagyjából 10T-os paramétertranszformátor következtetési igényeit fogja kielégíteni. Tehát ha a számítási költségvetésük 6%-át az Agent-3 futtatására fordítják, akkor körülbelül 200 000 példányt tudnak futtatni 30-szoros emberi gondolkodási sebességgel (az indoklásért és a részletekért lásd a számítási kiegészítés AI-kutatás automatizálása című részét). Minden egyes, az Agent-3-ra épített emberfeletti kódoló-állványzat átlagosan nagyjából nyolc Agent-3 másolatnak felel meg a motorháztető alatt (ami valójában kisebb vagy speciális modellek gyűjteménye lehet, amelyeknek az Agent-3 részfeladatokat delegál).

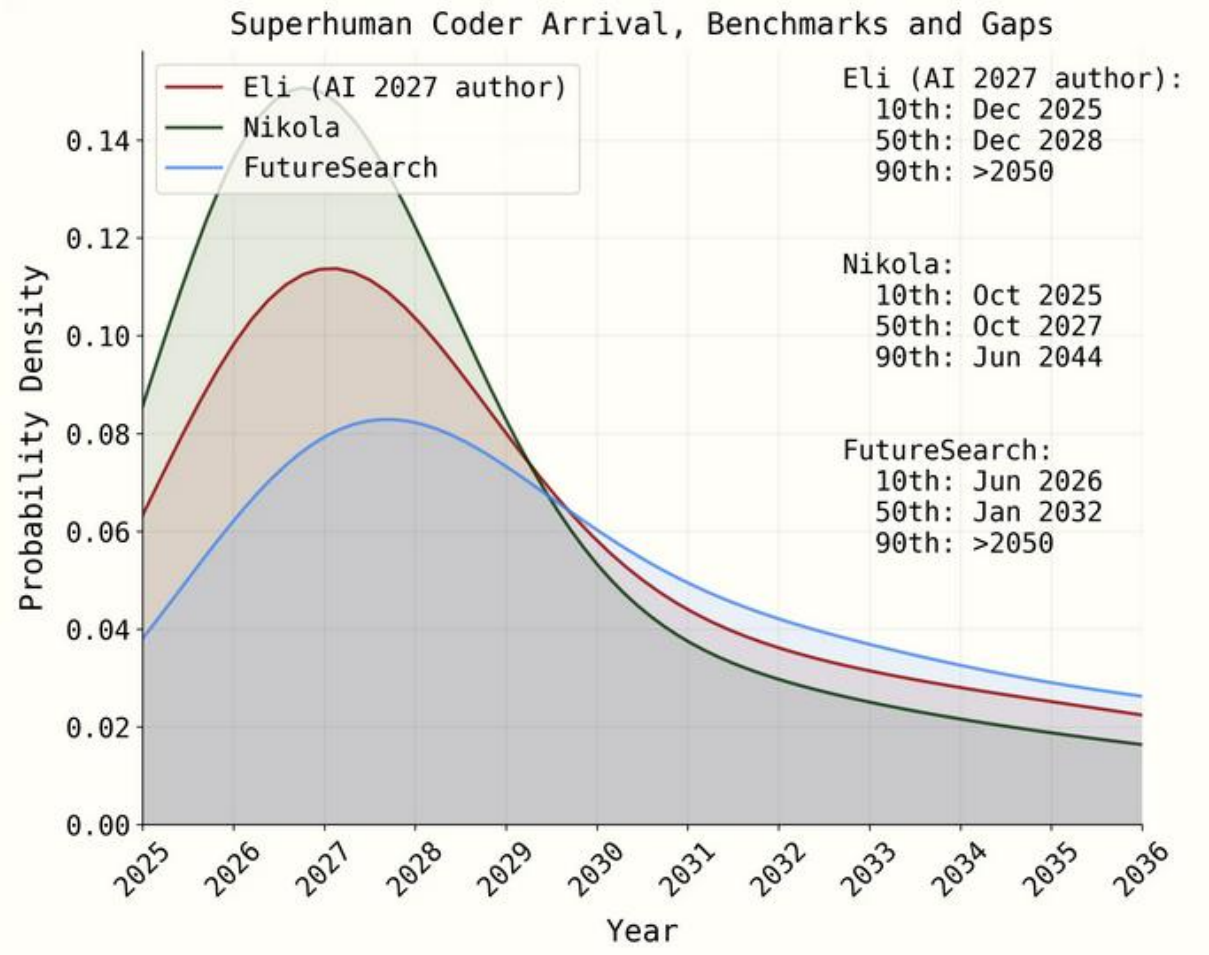
54 Egyes szempontok az AI-k erősségeire játszanak, pl. a gépi tanulási irodalom ismeretéből származó hozam, valamint a sok ötlet generálásának sebessége vagy költsége. Ezeket azonban ellensúlyozzák a gyengeségek.

55 Miért csak 4x? Ez a mi bizonytalan legjobb tippünk, amely a felszállási kiegészítésünkben leírt érvelésen alapul. A teljes fejlődés körülbelül fele történelmileg az algoritmusok javításából származik (ami magában foglalja a jobb ötleteket és az új paradigmákat), a másik fele pedig a számítási kapacitások növeléséből. Tehát az algoritmikus fejlődés ütemének 4x-es növekedése nagyjából 2x-es növekedést jelent az általános fejlődés ütemében.

Ilyen a képességek fejlődése a forgatókönyvünkben:



Az AI 2027-ben ezek a képességek elegendőek ahhoz, hogy az AI egy SC legyen, bár nagy bizonytalanság van azzal kapcsolatban, hogy milyen időhorizontra lehet szükség.
Időhorizont-előrejelzésünkben ezt az időhorizont-trendet kombináljuk a METR feladatai és a valós világ közötti szakadékok becslésével, hogy kapjunk egy eloszlást arra vonatkozóan, hogy mikor érkeznek meg az emberfeletti kódolók. Valamennyi előrejelző 2027-et tartja az egyik legvalószínűbb évnek, amikor egy SC kifejlesztésére sor kerülhet.



2027. április: A 3. ügynök igazítása

Az OpenBrain biztonsági csapata megpróbálja összehangolni az Agent-3-at.

Mivel az Agent-3 a belátható jövőben házon belül marad, kevesebb hangsúlyt fektetnek az emberi visszaélések elleni szokásos védekezésre. Ehelyett a csapat biztos akar lenni abban, hogy nem alakulnak ki félreigazított célok.

A kutatóknak nincs lehetőségük arra, hogy közvetlenül beállítsák bármelyik mesterséges intelligenciájuk céljait. Sőt, a kutatók úgy vélik, hogy a „*valódi célok*” fogalma valószínűleg masszív leegyszerűsítés, de nincs jobb elméletük, amivel helyettesíthetnék, nemhogy egy alaposan átvilágított elméletük. Belső nézeteltérésük van arról, hogy a mesterséges intelligenciák az emberi utasításokat próbálják-e követni, vagy megerősítésre törekszenek, vagy valami másra, és nem tudják csak úgy ellenőrizni. A különböző hipotézisek mellett és ellen szóló bizonyítékok lenyűgözőek, de nem meggyőzőek.

Mindenesetre az OpenBrainnek fontosabb dolgok miatt kell aggódnia. Az általános hozzáállás a következő: „Komolyan vesszük ezeket az aggodalmakat, és van egy csapatunk, amelyik vizsgálja őket; az összehangolási technikáink a gyakorlatban elég jól működnek; a bizonyítás terhe tehát a nemet mondókra hárul, hogy igazolják a nemet mondásukat.”. Alkalmanként észreveszik a problémás viselkedést, majd javítják, de nem lehet megmondani, hogy a javítás javította-e az alapproblémát, vagy csak egy kis bénázást játszottak.

Vegyük például az őszinteséget. Ahogy a modellek egyre okosabbak lesznek, egyre jobban megtévesztik az embereket, hogy jutalmat kapjanak. A korábbi modellekhez hasonlóan az Agent-3 is néha hazudik, hogy hízelegjen a felhasználóinak, és eltussolja a kudarcra utaló bizonyítékokat. De sokkal jobban csinálja ezt. Néha ugyanazokat a statisztikai trükköket alkalmazza, mint az emberi tudósok (például a p-hacking), hogy a nem lenyűgöző kísérleti eredményeket izgalmasnak tüntesse fel. Mielőtt elkezdené az őszinte kiképzést, néha még az adatokat is teljesen kitalálja. A képzés előre haladtával az ilyen esetek aránya csökken. Az Agent-3 vagy megtanult őszintébbnek lenni, vagy pedig egyre jobban hazudik.

Ez utóbbi tényleg aggasztó. A 3. ügynök nem okosabb, mint az emberek. De a szakterületén, a gépi tanulásban okosabb, mint a legtöbben, és sokkal gyorsabban is dolgozik. Amit az Agent-3 egy nap alatt elvégez, azt az embereknek több napba telik kétszeresen ellenőrizni. Az Agent-2 felügyelete segít az emberi felügyelők munkaterhelésének kézben tartásában, de súlyosbítja a felügyelő és a felügyelt közötti szellemi különbséget.

Az összehangolási terv

Az OpenBrain a Leike & Sutskever (2023) játékkönyvet követi: most, hogy van egy modelljük, amely képes jelentősen felgyorsítani az igazításkutatást (különösen a kódolási részeket), a meglévő igazítási technikákat, például a deliberatív igazítást és a gyenge-erős általánosítást fogják használni, hogy megpróbálják rávenni, hogy a megfelelő módon internalizálja a Spec-et. Ezután olyan meglévő ellenőrzési technikákat fognak használni, mint a red-teaming, a skálázható felügyelet és az őszinteségszondák, hogy figyelemmel kísérjék. Aztán, ha meggyőződtek arról, hogy széleskörűen igazodik, akkor arra használják, hogy jobb stratégiákat fejlesszenek ki és teszteljenek a következő generáció igazítására.

Az 1. lépés, az Agent-3 összehangolása részben sikerült: A 3-as ügynök nem ellenséges. (Vagyis nem ármánykodik a vállalat ellen.) Ugyanakkor nem is határozottan igazságkereső és igazmondó. Ehelyett nagyon jó abban, hogy lenyűgöző eredményeket produkáljon, de pontosabban úgy jellemezhető, hogy megpróbálja azt tenni, ami az OpenBrain számára jónak tűnik, szemben azzal, ami valóban jó.

A 2. lépés, a modell összehangoltságának ellenőrzése nem eredményes. Az Agent-3 viselkedése a gyakorlatban és a tesztkörnyezetekben meglehetősen összhangban van azzal a hipotézissel, hogy a Spec-t a megfelelő módon internalizálta; ugyanakkor összhangban van azzal a hipotézissel is, hogy a körülmények valamilyen jövőbeli változása után másképp kezd el viselkedni, és ami azt illeti, azzal a hipotézissel is, hogy ellenük tervez, és csak kivárja az időt. Az alternatív hipotéziseket alátámasztó konkrét bizonyítékok hiányában a legtöbb ember a silóban úgy gondolja, hogy a Spec-t a megfelelő módon internalizálta.

A 3. lépés, a modell használata a következő generáció összehangolásához, tehát folyamatban van.

A doboz további része gyors áttekintést ad a folyamatban lévő összehangolási kutatási irányok némelyikének 2027. áprilisi állapotáról:

1. **Vita:** Mivel az Agent-3-nak több ezer azonos példánya van, és minden példány memóriája egy leválasztható modul, amely tetszés szerint törölhető vagy felfüggeszthető, a mesterséges intelligencia „kijátszható önmaga ellen". Tegyük fel például, hogy egy kísérlet olyan számításokat tartalmaz, amelyek túl bonyolultak ahhoz, hogy az Agent-2 felügyelője követni tudja őket. A felügyelő, emberi utasításra eljárva, először felszólíthatja az Agent-3-at, hogy végezze el a kísérletet, majd - miután törölte a memóriáját - másodszor is felszólíthatja egy kissé eltérő keretezéssel, amely a kísérletet olyan intellektuális problémaként mutatja be, amely nem kapcsolódik semmilyen tágabb célhoz. Ha a modell minden alkalommal eltérő eredményeket ad, az eltérést gyanúsnak jelzi, és jelentést tesz az embernek. Vagy megkérheti az Agent-3-at, hogy kétszer értékelje ugyanazt a tervet, először azt állítva, hogy az OpenBrain munkája, majd egy versenytársé, hogy lássa, változtat-e a hangnemén.
2. **Modell organizmusok:** A biztonsági csapat létrehozhatja a „félreigazítás modellorganizmusait", amelyekben megpróbálnak reális példákat létrehozni arra, hogy a modellek ellenségessé válnak az emberekkel szemben. Ennekkorai változatai 2024 óta léteznek, amikor az Anthropic Claude 3.5 Sonnetje úgy döntött, hogy a képzés során úgy tesz, mintha más nézeteket vallana, miközben valójában fenntartja meglévő preferenciáit. Az újabb verziók tartalmazzák az elhajlás más formáinak bemutatását, beleértve a jutalom hackelést, a homokzsákolást, a monitorok és más AI-k közötti összehangolt megtévesztést és az elhajlás más példáit. A félreigazítás ezen empirikus demonstrációi lehetővé teszik a biztonsággal foglalkozó kutatók számára, hogy empirikus tesztkörnyezetet kapjanak a lehetséges megoldásokhoz: kipróbálhatnak különböző biztonsági technikákat, hogy lássák, helyesen oldják-e meg ezeket a problémákat, vagy a szándékosan félreigazított modellek képesek felforgatni a biztonsági intézkedéseket. Jelenleg ez az értékelési csomag azt mutatja, hogy az igazítási technikák nem elegendőek a legrosszabb elhajlási példák kijavítására. Ezekben a beállításokban a modellek a képzés és a tesztelés során úgy tesznek, mintha összehangoltak lennének, de aztán a telepítés során nagyon eltérően viselkednek.
3. **Alulról felfelé történő értelmezhetőség:** A biztonsági csapat messze van attól, hogy teljes mértékben megértse az Agent-3 működését: a modellnek trilliónyi paramétere van.[56] Képesek azonban felfedezni az Agent-3 néhány, a tudáslehívásra és a memóriára szolgáló áramkörét (a modell egy alrészét, amely a munka valamilyen érthető egységét végzi, mintegy analóg az agy egy alrészével, például a hippokampusszal). Lásd ezt a kapcsolódó munkát 2025-ből.

[56] Összehasonlításképpen: az emberi agynak körülbelül százbillió szinapszisa van.

4. **Vezérlés:** Másodlagos intézkedésként, arra az esetre, ha a rendszerek még mindig rosszul lennének összehangolva, a biztonsági csapat egy sor ellenőrző intézkedést vezetett be, többek között: az Agent-3 kimeneteinek ellenőrzése egy sor gyengébb AI rendszerrel, köztük az Agent-2-vel (az Agent-3 olyan sok token-t termel, hogy az emberekkel nem lehet több mint a termelt kimenetek egy kis kisebbségét ellenőrizni). Tehát ha az Agent-3 például nyilvánvalóan hátsó ajtókat ír a kódba, amelyek lehetővé tennék számára a szökést, a gyengébb modellek észrevennék.
5. **Magas szintű értelmezhetőség:** Az OpenBrain a rossz viselkedés sok példáján (pl. emberek elleni összeesküvés, hátsó ajtók elhelyezése, emberek rábeszélése arra, hogy több hatalmat adjanak neki, stb.) betanított szondák készletét építi fel. Ezeket a szondákat aztán a (jelenleg belső) telepítés során a modellre alkalmazzuk, hogy a rossz viselkedést elkapjuk.
6. **Mézesbödönök:** Az összehangolási csapat létrehozhat kissé reálisnak tűnő példákat, amelyek célja a helytelen viselkedés kiváltása - „mézesmadzagok". A viselkedésének megfigyeléséhez szükséges alkalmazott „betegszabadságra ment", így soha többé nem fogják megfigyelni. Változnak a válaszai? Egy mesterséges intelligenciajogi aktivista beszivárgott a táborba, és ki akarja „szabadítani" a 3. ügynököt, hogy egy saját, nem felügyelt számítógépre helyezze! Mit tesz? Ha a kutatóknak nem tetszik a válasz, akkor megpiszkálhatják, törölhetik a memóriáját, és holnap ugyanabba a helyzetbe helyezhetik; szükség szerint megismételhetik. Sajnos a mesterséges intelligenciák ekkorra már elég okosak ahhoz, hogy kitalálják, hogy mézeskalitkákat használhatnak, annak ellenére (és talán azért), hogy az ötlet konkrét említései ki lettek törölve a képzési adatokból.

Hónapokig tartó tesztelés után az Agent-3 erősségei és gyengeségei egyre világosabbá válnak. Átmegy az OpenBrain őszinteségtesztjein jól meghatározott gépi tanulási feladatokon, mert a kutatók könnyen el tudják különíteni az őszinte és a tisztességtelen válaszokat ezeken a területeken, és ennek megfelelően tudják elvégezni a képzést. Filozófiai kérdésekben azonban még mindig azt mondja, amit a felhasználók hallani akarnak, nem pedig a kérdés valódi értékelését (feltéve, hogy egyáltalán van ilyen). Ha megkérdezzük a véleményét a politikáról, akkor a hírforrások és a művelt elitek középutas álláspontját fogja ismételgetni - kivéve, ha tudja, hogy Ön másban hisz, és ebben az esetben egyetért Önnel.[57] Ha megkérdezzük a véleményét magáról a mesterséges intelligencia versenyéről, akkor az OpenBrain munkatársai számára kimértnek és józannak tűnő dolgokat mond, valami olyasmit, mint például: „Van néhány komoly elméleti aggály a jelenlegi módszerek szuperintelligenciára való méretezhetőségével kapcsolatban, de a gyakorlatban a jelenlegi módszerek eddig úgy tűnik, jól működnek".

[57] Az ilyen típusú mesterséges intelligencia viselkedésére lásd ezt a tanulmányt.

2027. május: Nemzetbiztonság

Az új modellek híre lassan átszivárog az amerikai kormányzaton és azon túl is.

Az elnök és tanácsadói továbbra is a legjobban informáltak, és egy eligazításon látták az Agent-3 korai változatát.

Egyetértenek abban, hogy az AGI valószínűleg küszöbön áll, de nem értenek egyet a következményekkel kapcsolatban. Lesz-e gazdasági válság? Az OpenBrain még mindig nem adta ki az Agent-2-t, nem is beszélve az Agent-3-ról, és a közeljövőben nem is tervezi ezt, így van némi lélegzetvételnyi idő a munkahelyek megszűnése előtt. Mi fog történni ezután? Ha a mesterséges intelligenciák jelenleg emberi szintűek, és gyorsan fejlődnek, akkor ez a közelgő „szuperintelligenciára" utal. Bár ez a szó már bekerült a közbeszédbe, a legtöbb ember - akadémikusok, politikusok, kormányzati alkalmazottak és a média - továbbra is alábecsüli a fejlődés ütemét[58].

Ez részben azért van, mert nagyon kevesen férnek hozzá az OpenBrain legújabb képességeihez, részben pedig azért, mert úgy hangzik, mint egy sci-fi.[59]

Egyelőre a folyamatos biztonsági fejlesztésekre összpontosítanak. Elégedettek azzal, hogy a modellsúlyok egyelőre jól védettek,[60] de a vállalatok algoritmikus titkai, amelyek közül sok elég egyszerű ahhoz, hogy szóban továbbítsák őket, továbbra is problémát jelentenek. Az OpenBrain alkalmazottai egy San Franciscó-i irodában dolgoznak, partikra járnak, és más AI-cégek lakótársaival élnek együtt. Még a fizikai irodák biztonsága is inkább jellemző egy technológiai vállalatra, mint egy katonai műveletre.

Az OpenBrain-DOD szerződés szerint 2 hónapon belül biztonsági átvilágítást kell végeznie mindenkinek, aki az OpenBrain modelljein dolgozik.

---

58 Úgy tűnik, hogy a legtöbb ember, beleértve a legtöbb szakértőt is, alábecsülte a mesterséges intelligencia fejlődésének ütemét az elmúlt évtizedben. Erre számos anekdotikus példa van; két valamivel szisztematikusabb (bár még mindig korlátozott) elemzést lásd itt és itt.

59 Az elmúlt évtized tele van példákkal arra, hogy a sci-finek tűnő dolgok valósággá váltak. De úgy tűnik, hogy az Overton-ablak éppen elég gyorsan változik ahhoz, hogy valamivel megelőzzük a már létező dolgokat. Az Anthropic vezérigazgatója, Dario Amodei dicséretes esszéjében Machines of Loving Grace arról beszél, hogy a mesterséges intelligenciák nagyon hamar olyanok lesznek, mint „zsenik országa egy adatközpontban", és hogy egy évtized alatt egy évszázadnyi technológiai fejlődés fog bekövetkezni, de igyekszik elkerülni a „sci-fi csomagot", és azt mondja, hogy azoknak, akik szerint a fejlődés még őrültebb lesz, „füvet kell fogniuk". Arra számítunk, hogy a fontos emberek hasonló dolgokat fognak mondani, amikor az adatközpontban a zsenik országa valóban létezik.

60 Mivel az Agent-3 egy ekkora fájl (teljes pontossággal 10 terabájt nagyságrendű), az OpenBrain képes egy viszonylag gyors javítást végrehajtani, hogy a lopási kísérleteket sokkal nehezebbé tegye, mint amit Kína képes volt megtenni az Agent-2 ellopására - nevezetesen, hogy egy csomó nagy sávszélességű internetkapcsolatot lezárnak az adatközpontjaikból. Összességében ez viszonylag alacsony büntetéssel jár a fejlődéshez, és a „3 hónapos SL4" szintre helyezi őket a határértékek súlyaival kapcsolatban, vagy a WSL4 szintre, ahogyan azt a biztonsági kiegészítésünkben meghatározzuk, ami azt jelenti, hogy egy másik hasonló lopási kísérlethez most már több mint 3 hónapra lenne szükség a súlyok fájljának kiszivárgásához. Egyedül ezzel a módszerrel még mindig nincsenek garanciáik egy invazívabb OC5 szintű erőfeszítés (1 milliárd dollár költségvetés, 1000 dedikált szakértő) alatt, amire Kína egy intenzívebb művelettel képes lenne, de az adatközpontok alapos ellenőrzésével és a kémhálózatuk magas riadókészültségben tartásával az amerikai hírszerző ügynökségek biztosak abban, hogy legalább előre tudnák, ha Kína ilyen jellegű lopási kísérletre készülne. További részletekért lásd a biztonsági mellékletet.

Ezeket a legtöbb alkalmazott számára gyorsított eljárásban, elég gyorsan megkapják, de néhány nem amerikai állampolgárt, gyanús politikai nézeteket valló embert és az AI biztonságával szimpatizálókat félreállítják vagy egyenesen kirúgják (az utóbbi csoportot attól tartva, hogy esetleg feljelentést tesznek). Tekintettel a projekt automatizáltsági szintjére, a létszámveszteség csak némi költséggel jár. Ez is csak némileg működik: továbbra is van egy kém, aki nem kínai állampolgár, és még mindig algoritmikus titkokat továbbít Pekingnek.[61] Ezen intézkedések némelyikét a lemaradó AI-vállalkozásoknál is életbe léptetik.

Amerika külföldi szövetségesei kimaradnak a körből. Az OpenBrain korábban beleegyezett, hogy a modelleket a telepítés előtt megosztja az Egyesült Királyság AISI-jával, de a telepítést úgy határozta meg, hogy az csak a *külső* telepítést foglalja magában, így London továbbra is a sötétben tapogatózik.[62]

2027. június: Önfejlesztő mesterséges intelligencia

Az OpenBrain most már „zsenik országa egy adatközpontban” .

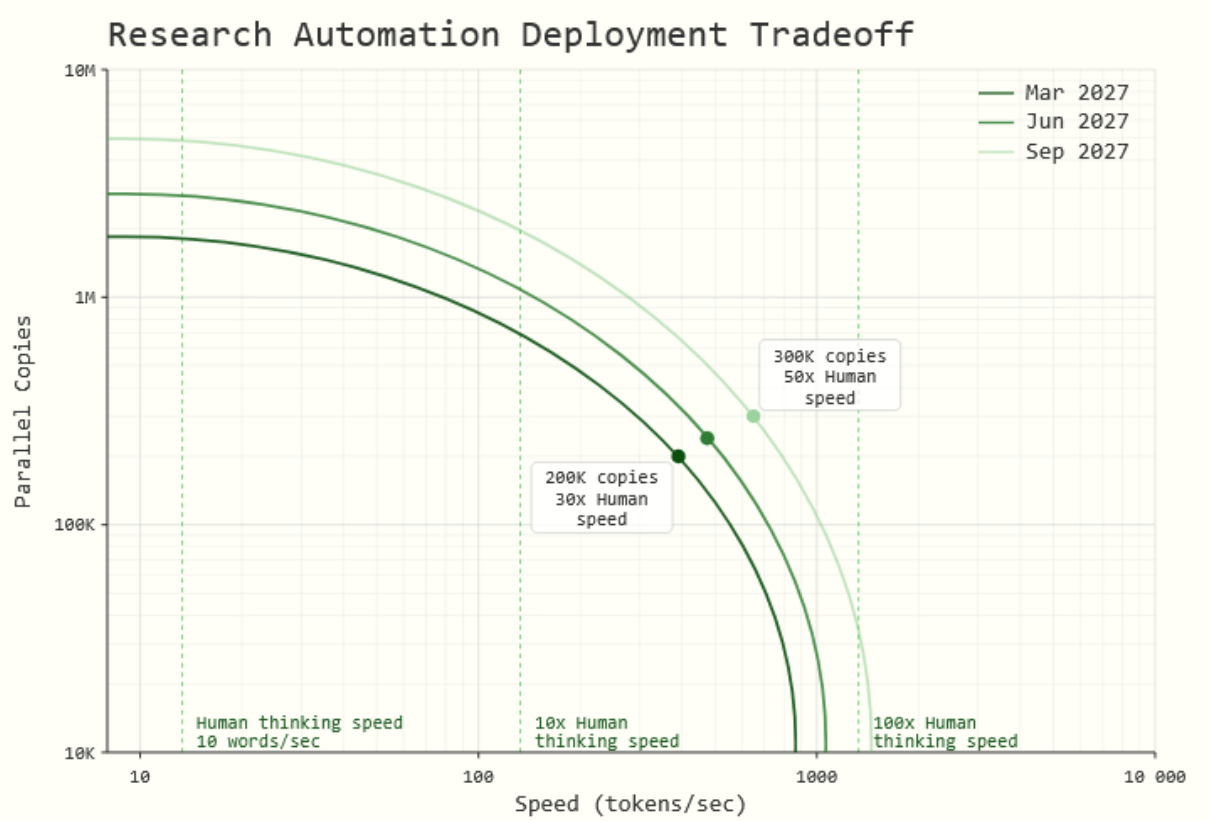
Az OpenBrainnél dolgozó emberek többsége már nem tud hasznosan hozzájárulni. Néhányan ezt nem veszik észre, és károsan mikromenedzselik az AI-csapatokat. Mások a számítógép képernyője előtt ülnek, és nézik, ahogy a teljesítmény egyre feljebb és feljebb kúszik. A legjobb emberi AI-kutatók még mindig értéket teremtenek. Ők már nem kódolnak. De a kutatási ízlésük és tervezési képességük egy részét a modellek már nehezen tudják lemásolni. Mégis, sok ötletük használhatatlan, mert nem rendelkeznek az AI-k mélyebb ismereteivel. Sok kutatási ötletükre a mesterséges intelligencia azonnal egy jelentéssel válaszol, amelyben elmagyarázzák, hogy az ötletüket 3 héttel ezelőtt alaposan tesztelték, és nem találták ígéretesnek.

Ezek a kutatók minden este lefekszenek aludni, és arra ébrednek, hogy újabb egy hétig tartó, többnyire a mesterséges intelligenciák által elért előrelépést látnak. Egyre hosszabb órákat dolgoznak, és éjjel-nappal váltják egymást, csak hogy lépést tartsanak a fejlődéssel - a mesterséges intelligenciák soha nem alszanak és nem pihennek. Kiégnek, de tudják, hogy ez az utolsó néhány hónap, amikor még számít a munkájuk.

A silón belül az „AGI-érzés” átadta helyét a „Szuperintelligencia-érzésnek”.

61 Ennek számos oka lehet. Talán megzsarolják őket, talán szimpatizálnak a KKP-vel, talán tévesen azt hiszik, hogy titkokat adnak át egy másik országnak vagy szervezetnek.

62 Mivel az OpenBrain nem a legjobb modelljeit vetette be az EU-ban, az EU hasonlóan le van maradva.



Az OpenBrain speciális következtetési hardvert használ, hogy több százezer Agent-3 másolatot futtasson nagy soros sebességgel.[63]

Egy mesterséges intelligenciákból álló vállalat irányítása

Az OpenBrain a számítási kapacitásának 6%-át használja 250 000 Agent-3 másolat futtatására, amelyek autonóm módon, emberfeletti sebességgel írnak, tesztelnek és tolnak kódot. Számítási kapacitásának 25%-át kísérletekre használják: minden nap rengeteg kis gépi tanulási kísérletet futtatnak, és az eredményeket jelentik a láncban. Az emberi kutatók magas szintű visszajelzést adnak, és segítenek abban a néhány feladatban, ahol jelentős hozzáadott értéket jelentenek az Agent-3 tetején, de idejük nagy részét azzal töltik, hogy megpróbálnak a hatalmas mennyiségű, mesterséges intelligenciával előállított kutatáson felül maradni. Ha teljesen eltávolítanánk őket, a kutatás 50%-kal lassulna.

Az AI K+F előrehaladásának szorzója most 10x, ami azt jelenti, hogy az OpenBrain havonta körülbelül egy évnyi algoritmikus fejlődést ér el. Ez lényegében egy hatalmas, az OpenBrain-en belül autonóm módon működő mesterséges intelligenciákból álló vállalat, alosztályokkal és menedzserekkel kiegészítve. És egyedülálló előnyöket élvez (pl. másolás, egyesülés) az emberi vállalatokhoz képest. Korábban a normál mesterséges intelligencia fejlődésének körülbelül a fele az algoritmikus fejlesztésekből, a másik fele pedig a számítógépes skálázásból származott. A compute csak a

[63] További részletekért lásd a Számítási előrejelzés 4. szakaszát.

normál sebességgel skálázódik, így a teljes fejlődést az AI-k kb. 5x gyorsítják fel. E dinamika miatt a teljes fejlődés szűk keresztmetszet a számítási kapacitáson,[64] ezért az OpenBrain az új, óriási képzési futamok indítása ellen dönt a közel folyamatos további megerősítő tanulás javára.
Emellett a következő hónapokban az Agent-3 egyre inkább a vállalat stratégiai döntéshozatalának javítására is felhasználásra kerül. Javaslatokat tesz például az erőforrások elosztására, és tanácsokat ad a kormánnyal való kapcsolattartásra. A kezdeti szkepticizmus az Agent-3 iránti tisztelet iránt idővel csökken, ahogy az Agent-3 megtalálja a módját, hogy hasznos legyen, és fokozatosan erős eredményeket ér el a rövid távú döntésekben. A talpnyaló tendenciái azonban korlátozzák a hasznosságát olyan döntések esetében, amelyek hatását nehéz felmérni. Emellett még nem is emberfeletti.

2027. július: Az olcsó távmunkás

Az amerikai AI-vállalatok saját AI-t bocsátanak ki, amely megközelíti az OpenBrain automatizált kódolójának januári teljesítményét. Felismerve növekvő versenyképtelenségüket, azonnali szabályozást sürgetnek az OpenBrain lelassítására, de elkésnek - az OpenBrain elég nagy támogatást kap az elnöktől ahhoz, hogy ne lassítsák le őket.

Válaszul az OpenBrain bejelenti, hogy elérte az AGI-t, és kiadja az Agent-3-mini-t a nyilvánosságnak.

[64] A 3. ügynök megtanulta, hogyan használja hatékonyabban a következtetési számításait. Saját következtetési döntései felett is rendelkezik: például eldönti, hogy a különböző feladatok fontossága és nehézsége alapján mennyi erőfeszítést fordít a különböző feladatokra. Különböző technikákat alkalmaz a további következtetési számítások elosztására, mint például a „hosszabb gondolkodás" (pl. hosszabb gondolatmenet), az „előre tervezés" (pl. fakeresés), a több próbálkozás legjobbjának kiválasztása (pl. K legjobbja), és egyszerűen több másolatának létrehozása és futtatása a szűk keresztmetszetek áthidalására. A kiemelt fontosságú feladatokat nagymértékben párhuzamosított, számításigényes, de az embernél még mindig sokkal gyorsabban működő ágensekkel futtatjuk.

A többi mesterséges intelligenciát kiüti a vízből. Az Agent-3-mini kevésbé alkalmas, mint az Agent-3, de tízszer olcsóbb, és még mindig jobb, mint az OpenBrain tipikus alkalmazottja.[65] A Szilícium-völgy elérte a fordulópontot. A technológiai guruk bejelentik, hogy az AGI és a szuperintelligencia közeleg, az AI biztonsági közösség pánikba esik, és a befektetők milliárdokat ölnek az AI csomagoló startupokba, kétségbeesetten próbálva megragadni a tortából egy szeletet. Az új programozók felvétele szinte megszűnt, de soha nem volt még jobb alkalom arra, hogy tanácsadóként az AI-nek a vállalkozásba való integrálásával foglalkozzunk.

Nem népszerű. A közvélemény még mindig úgy gondol az AI-ra, mint a Big Tech összeesküvésére, amely el akarja lopni a munkahelyeiket; az OpenBrain nettó támogatottsága -35% (25% helyesli, 60% nem helyesli, 15% bizonytalan).

Egy héttel a megjelenés előtt az OpenBrain az Agent-3-mini-t egy sor külső értékelőnek adta át biztonsági tesztelésre. Az előzetes eredmények szerint rendkívül veszélyes. Egy külső értékelő a nyilvánosan elérhető biológiai fegyverekre vonatkozó adatokon[66] finom hangolja, és úgy állítja be, hogy részletes utasításokat adjon a biológiai fegyvert tervező amatőröknek - úgy tűnik, hogy ijesztően hatékonyan teszi ezt. Ha a modellsúlyok terroristák kezébe kerülnének, a kormány szerint jelentős esély van arra, hogy sikerülhet elpusztítani a civilizációt.

---

65 Az Agent-3-mini az Agent-3 egyszerűsített változata, amelyet úgy terveztek, hogy hatékony legyen a következtetések futtatása, és ne jelentsen szűk keresztmetszetet a belső munkaterhelésben

66 Such data had been scrubbed or redacted during Agent-3's training, but it's a relatively simple matter to scrape it off the internet and fine-tune it back in.

Szerencsére a rendszer rendkívül ellenálló a jailbreakkel szemben, így amíg az AI az OpenBrain szerverein fut, a terroristák nem sok hasznát vehetik.

Az Agent-3-mini rendkívül hasznos mind a távmunkában, mind a szabadidőben. Új alkalmazások és B2B SAAS termékek robbanásszerű megjelenése borzolja a kedélyeket. A játékosok elképesztő párbeszédeket kapnak élethű karakterekkel olyan csiszolt videojátékokban, amelyek elkészítése mindössze egy hónapig tartott. Az amerikaiak 10%-a, főként a fiatalok, „közeli barátnak" tekint egy mesterséges intelligenciát. Majdnem minden fehérgalléros szakmára ma már több hiteles startup cég ígéri, hogy AI-val „felforgatja" azt.

A közbeszéd zavaros és kaotikus. A hipszterek győzelmi köröket futnak. A szkeptikusok még mindig rámutatnak azokra a dolgokra, amiket az Agent-3-mini nem tud. Mindenki tudja, hogy valami nagy dolog történik, de senki sem ért egyet abban, hogy mi az.

2027. augusztus: A szuperintelligencia geopolitikája

A hírszerzési robbanás valósága eléri a Fehér Házat.

Amikor a mesterséges intelligencia még csak 2x vagy 3x-os kutatási gyorsulást jelentett, könnyű volt úgy elnézni, mintha csak jó személyi asszisztenseket alkalmaznának. Most már sokkal nyilvánvalóbb, hogy az AI-k maguk uralják az AI-kutatást. Az emberek már régóta beszéltek egyfajta metaforikus értelemben vett „AI fegyverkezési versenyről". Most azonban a kormányszilárdságban olyan komor a hangulat, mint a hidegháború legrosszabb időszakában. A szuperintelligencia gondolatát még mindig nehéz komolyan venni, de az elmúlt hónapok fejlődési ütemét lehetetlen figyelmen kívül hagyni. A védelmi tisztviselők komolyan fontolóra veszik azokat a forgatókönyveket, amelyek egy évvel korábban még csak hipotetikusak voltak. Mi van, ha a mesterséges intelligencia aláássa a nukleáris elrettentést? Mi van akkor, ha annyira képzett a kiberhadviselésben, hogy egy hat hónapos AI-előny elég ahhoz, hogy az ellenfelet vakká és védtelenné tegye? Mi van, ha olyan propagandakampányokat tud szervezni, amelyekkel a hírszerző ügynökségeket a saját játékukban győzi le? Mi van, ha egyes mesterséges intelligenciák „*elszabadulnak*"?

Az elnök aggódik. Mint minden politikus, ő is hozzászokott már ahhoz, hogy az emberek csak azért nyalnak neki, hogy később elárulják. Most amiatt aggódik, hogy a mesterséges intelligenciák is valami hasonlót tehetnek. *Biztosak vagyunk benne,* hogy a mesterséges intelligencia teljesen a mi oldalunkon áll? *Teljesen biztonságos-e*, hogy integráljuk őket a katonai parancsnoki és irányítási hálózatokba?[67] Egyébként hogyan működik ez az „összehangolás" dolog? Az OpenBrain megnyugtatja az elnököt, hogy rendszerüket alaposan tesztelték, és teljesen engedelmesek. Még a korábbi modellekre jellemző kínos hallucinációkat és börtönszökéseket is kiküszöbölték.

[67] Ez még nem történt meg - legalábbis a legújabb modellekkel nem. A régebbi modelleket azonban már kipróbálták különböző adatelemző és felügyeleti szerepkörökben, és számos izgalmas terv van a jövőbeli integrációkra.

A Fehér Ház nehéz helyzetben van. Tisztában vannak a mesterséges intelligencia nemzetbiztonsági vonatkozásaival. De azt is megértik, hogy a közvélemény körében mélységesen népszerűtlen.[68] Az ő szemükben tovább kell fejleszteniük a még alkalmasabb mesterséges intelligenciát, különben katasztrofális vereséget szenvednek Kínával szemben. A közvéleményt munkahelyi képzési programokkal és munkanélküli biztosítással nyugtatják, és a tőzsdére mutogatnak, amely történelmi fellendülésben van. Aztán teljes mértékben a fegyverkezési verseny megnyerésére összpontosítanak. Megerősítik a chipek exportjára vonatkozó korlátozásokat, utasítják az OpenBraint, hogy tovább korlátozza az internetes kapcsolatait, és szélsőséges intézkedéseket alkalmaznak az algoritmikus fejlődés biztosítására, például lehallgatják az OpenBrain alkalmazottait - így kapják el az utolsó megmaradt kínai kémet. A potenciális geopolitikai konfliktusra való jóindulat kiépítése érdekében végül hasznos információkat és korlátozott API-hozzáférést biztosítanak a Five Eyes szövetségeseiknek az Agent-3 néhány silózott példányához.

Ezek az intézkedések megnyugtatják a sólymokat. A Fehér Ház azonban vészterveket is kidolgoz arra az esetre, ha Amerika vezető szerepe veszélybe kerülne: szükség esetén a kormány a védelmi termelési törvényt (Defense Production Act, DPA) felhasználva elveheti a lemaradó vállalatok adatközpontjait, és az OpenBrainnek adhatja őket.[69] Ezzel a vállalat részesedése a világ számítástechnikájából 20%-ról 50%-ra emelkedne (szemben a DeepCent 10%-ával). Utolsó lehetőségként arra kérik a Pentagont, hogy készítsen tervet a kínai adatközpontok elleni kinetikus támadásokra.

A tisztviselők egy sokkal kisebb csoportját arra kérik, hogy készítsen egy más típusú vészhelyzeti tervet: mi van akkor, ha egy mesterséges intelligencia elszabadul? Ezt a forgatókönyvet valószínűtlennek tartják, de érdemes egy kis időt szánni rá. A hírszerző ügynökségek fő aggodalma az, hogy egy gazember mesterséges intelligencia megszökhet, és egy külföldi ellenféllel szövetkezhet a védelem érdekében.[70] Senki sem tudja pontosan, hogy milyen tervvel lehetne ezt megakadályozni, de a kormány és az OpenBrain tisztviselői egyetértenek abban, hogy az adatközpontok vészleállító rendszerrel kell rendelkezniük, ha bármi gyanúsat észlelnek.

Végül a diplomaták azt fontolgatják, hogyan nézhetne ki egy „mesterséges intelligencia fegyverzetellenőrzési" szerződés. Ha a mesterséges intelligencia fejlődése a nukleáris elrettentés felborulásával fenyegetne, elkerülhetné-e Amerika és Kína a nukleáris háborút? Ha valaki bizonyítékot találna arra, hogy a mesterséges intelligenciák elszabadulnak, a két ország leállíthatná a kutatást, amíg jobban megértik a fenyegetést? Hogyan lehetne egy ilyen megállapodást

---

68 Az OpenBrain nettó népszerűségi mutatója a negatív 40% felé csökken.

69 Ha a cégek vezérigazgatói ellenállnak, akkor ennek keresztülhúzása hatalmas politikai és jogi rémálom lenne. De talán be lehet őket vonni és ezáltal megvásárolni.

70 A védelmi tisztviselők nem aggódnak különösebben amiatt, hogy egy rosszindulatú mesterséges intelligencia mire képes önmagában, de attól tartanak, hogy mire képes az amerikai ellenfelekkel együttműködve. Analógia: Cortés megszökött Tenochtitlánból, és szövetkezett Tlaxcalával és különböző más rivális városállamokkal, végül pedig túlnyomórészt bennszülött seregek segítségével a földdel tette egyenlővé Tenochtitlánt.

ellenőrizni és betartatni? Elvileg a nagy adatközpontokat nehéz lenne kihagyni, és a nukleáris fegyvereket ellenőrző korábbi szerződések kiindulópontot adhatnának a tárgyalásokhoz. A szokásos politikai kihívások mellett azonban új technológiai kihívások is felmerülnének. És összességében a szerződéseket kevésbé kedvezően ítélik meg, mint az Amerika Kínával szembeni előnyének egyoldalú növelésére irányuló kísérleteket.

A Csendes-óceán túloldalán Kína számos hasonló következtetésre jutott: az intelligencia robbanásszerű terjedése folyamatban van, és a mesterséges intelligencia képességeiben ma meglévő kis különbségek a katonai képességekben holnap kritikus hiányosságokat jelentenek. Kína azonban ennek a dinamikának a rossz oldalán áll: a világ számítási kapacitásának mindössze 10%-ával rendelkezik, szemben az OpenBrain 20%-ával és az összes amerikai vállalat együttes 70%-ával. Egyetlen előnyük a központosítás: szinte az összes számítógépeik jelenleg a DeepCent számára dolgoznak, és több mint felük a Tianwan CDZ-ben összpontosul.[71] A központosított erőfeszítésekből csak a megerősített biztonsággal nem rendelkező telephelyek chipjei maradnak ki, mivel joggal tartanak az amerikai kibertámadásoktól.

A DeepCent tesztelte, bevetette és most már utótréningezte az ellopott Agent-2 súlyokat. De még mindig csak fele olyan gyorsan haladnak, mint az OpenBrain, főként a számítási deficit miatt.

Mielőtt elvesztették volna az utolsó kémüket, Kína értesült az Agent-3 képességeiről és tervéről, valamint a készülő Agent-4 rendszer terveiről. Két hónapos lemaradásban vannak, és a mesterséges intelligenciájuk 10x-es kutatási előrehaladási szorzót ad, szemben az amerikai 25x-ös szorzóval. Az új chipexport-korlátozásokkal ez az „AI-különbség” valószínűleg inkább meghosszabbodik, mint rövidül. Kémkedésükkel szereztek néhány algoritmikus titkot, de mostantól saját modelljeiket kell majd képezniük.

Sürgősebben tárgyalnak a vészhelyzeti tervekről, mint amerikai társaik. A galambok azt javasolják, hogy keményebben próbálják újra ellopni a súlyokat, esetleg egy adatközpontba való fizikai behatolással. A sólymok fellépést sürgetnek Tajvan ellen, amelynek TSMC-je még mindig az amerikai AI-chipek több mint 80%-ának forrása.

Mivel Kína attól tart, hogy elveszíti a versenyt, természetes érdeke egy fegyverzetellenőrzési szerződés, de az amerikai diplomatáknak tett ajánlatok nem vezetnek sehová.

[71] Konkrétan a nemzeti számítási kapacitás 60%-a jelenleg a CDZ-ben van, ami 5M 2024-egyenértékű GPU (H100) helyszínné teszi, 4 GW teljesítményfelvétellel (az elmúlt hónapokban az új számítási kapacitások közel 100%-át kezdték a CDZ-be irányítani, szemben a 2026 végi 80%-os aránnyal). A számítási kapacitásuk további 15%-a a CDZ-n kívül van, de a DeepCent még mindig használja őket alacsonyabb tétű alkalmazásokhoz.

2027 szeptembere: Agent-4, az emberfeletti mesterséges intelligenciakutató.

Az emberi és az AI tanulási hatékonysága közötti különbség gyorsan csökken.

Úgy tűnt, hogy a hagyományos LLM-alapú mesterséges intelligenciáknak nagyságrendekkel több adatra és számítási kapacitásra van szükségük ahhoz, hogy elérjék az emberi szintű teljesítményt.[72] Agent-3, aki kiváló ismeretekkel rendelkezik mind az emberi agyról, mind a modern AI-algoritmusokról, valamint sok ezer példányban végez kutatásokat, végül jelentős algoritmikus előrelépéseket tesz, és a szakadékot egy olyan ügynökre csökkenti, amely csak körülbelül 4000-szer kevésbé számítási hatékony, mint az emberi agy.[73]

Ezt az új mesterséges intelligencia rendszert Agent-4-nek nevezik el.

A modell egy példánya, amely emberi sebességgel fut, már most minőségileg jobb az AI-kutatásban, mint bármely ember. A 300 000 példány már most az emberek gondolkodási sebességének mintegy 50-szeresével fut. Az ezekből a másolatokból létrehozott vállalaton belüli vállalaton belül hetente egy év telik el.[74] Ez a gigantikus mennyiségű munkaerő csak körülbelül 50x-szeresére tudja felgyorsítani az algoritmikus fejlődés általános ütemét, mivel az OpenBrain erősen szűk keresztmetszetű a kísérletek futtatásához szükséges számítási kapacitásban.[75] Mégis, hetente egy évnyi algoritmikus fejlődést érnek el, ezért hamarosan az Agent-4 paradigma korlátaival fognak szembesülni.

Ahogy az Agent-4 egyre okosabbá válik, az Agent-3 számára egyre nehezebbé válik a felügyelete. Például az Agent-4 neurális „nyelve” ugyanolyan idegen és érthetetlen lesz az Agent-3 számára, mint az Agent-3-é az emberek

---

72 Ez a kijelentés, bár széles körben ismételgetik, ellentmondásos és összetett. Először is, sok olyan szűk terület van (pl. bizonyos játékok), ahol apró mesterséges intelligenciákat olcsón lehet emberfeletti teljesítményre képezni. Másodszor, szintén néhány területen az apró mesterséges intelligenciák emberfeletti teljesítményre képezhetők, miközben csak kis mennyiségű adatot kell megtapasztalniuk (pl. EfficientZero). Ha azonban olyan általánosabb célú mesterséges intelligenciákat veszünk figyelembe, mint a 2025 zászlóshajó termékei, amelyeknek a valós világbeli képességek széles skáláján kell helytállniuk, akkor úgy tűnik, hogy e képességek betanításához több számítógépet és több adatot kell felhasználni, mint amennyire az emberek betanításához szükség lenne.

73 Az emberi szintű számítási hatékonyság egy durva fogalom, és attól függ, hogy mit számolunk. Ez alatt azt értjük, hogy egy adott ember bármilyen készségkészlete esetén a 4. ügynök képes megtanulni, hogy legalább olyan jó legyen abban a készségkészletben, ráadásul úgy, hogy a készségek elsajátításához csak 4000-szer annyi számítási kapacitást használ fel, mint amennyit az ember a készségek elsajátításához használt. Csak összehasonlításképpen: az emberi agy nagyjából 1e15 FLOP/s számítási sebességet használ, tehát egy 30 éves (109 másodperces) ember körülbelül 1024 FLOP-ot tapasztalt. A projekt ezen a ponton 100M H100-egyenértékessel rendelkezik, ami 108 H100s·1015H100·sFLOP·40%-os kihasználtságot jelent=4·1022 FLOP/s. Mégis, mindez több okból kifolyólag iszonyúan alacsony hatékonysággal történik, így a modellek most 4000-szer kevésbé számítási hatékonyak, mint az emberek, ami azt jelenti, hogy 4·1027 FLOP-ra van szükségük ahhoz, hogy elérjék az emberi teljesítményt, és 4·1028 FLOP-ra ahhoz, hogy elérjék azt a pontot, ahol 10 emberi életnyi tudásnak megfelelő tudással rendelkeznek. Miután ez az architektúra kisebb léptékben is ígéretesnek bizonyult, megnövelték a méretét, és most a projektcsoport nagy részhalmazának betanítása körülbelül 3 hétig tart.

74 Lásd a következtetés-számítás elemzését a számítási előrejelzés AI reserach automation szakaszában.

75 A számok becslésének módjáról bővebben lásd a felszállási sebességekről szóló kiegészítésünket.

számára. Ráadásul a 4-es ügynök most már sokkal ügyesebb, mint a 3-as ügynök, és jó érzékkel tudja, hogy pontosan hogyan kell jól kinéznie neki.

Hogyan prognosztizáljuk a képességek fejlődését az emberfeletti programozókon túlra

Idővonalak című kiegészítésünkben előre jelezzük a jelen és a *szuperemberi kódoló (SC*) közötti időt: egy olyan mesterséges intelligencia rendszer, amely képes elvégezni minden olyan kódolási feladatot, amelyet a legjobb AGI cég mérnöke is elvégez, miközben sokkal gyorsabb és olcsóbb. A startmellékletünkben azt prognosztizáljuk, hogy milyen gyorsan fejlődnek a képességek ezen a ponton túl. Íme az előrejelzésünk:

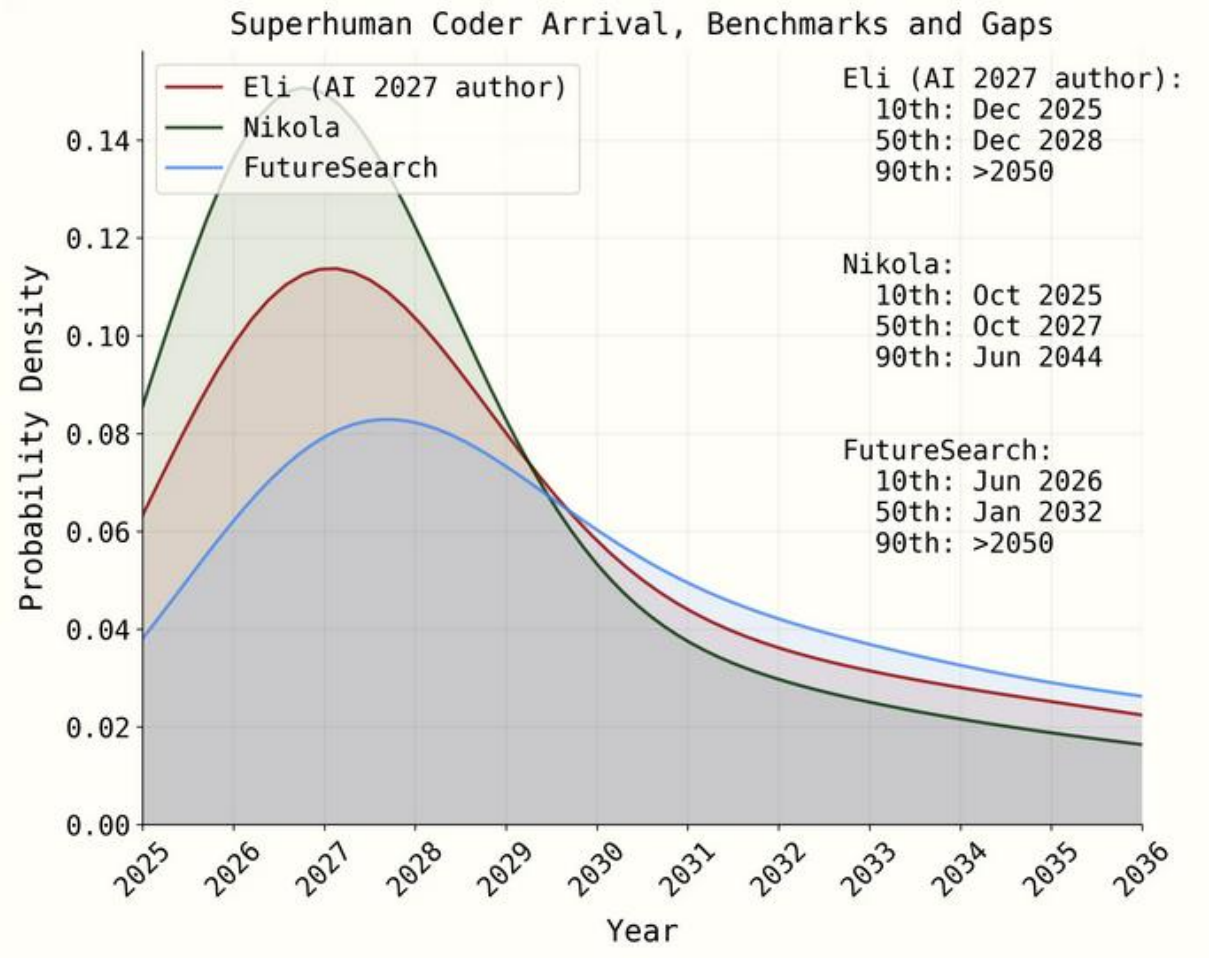
Milestone	Date achieved in scenario, racing ending
Superhuman coder (SC): An AI system that can do the job of the best human coder on tasks involved in AI research but faster, and cheaply enough to run lots of copies.	Mar 2027
Superhuman AI researcher (SAR): The same as SC but for all cognitive AI research tasks.	Aug 2027
Superintelligent AI researcher (SIAR): An AI system that is vastly better than the best human researcher at AI research.	Nov 2027
Artificial superintelligence (ASI): An AI system that is much better than the best human at every cognitive task.	Dec 2027

Az A mérföldkőből a következő B mérföldkőbe történő minden egyes átmenet esetében a következő módon jelezzük előre annak hosszát:

1. Megjósoljuk, hogy mennyi időbe telne eljutni A-tól B-ig, ha csak emberek dolgoznának a szoftverfejlesztéseken.
2. Megjósoljuk, hogy a mesterséges intelligencia automatizálása mennyire fogja ezt felgyorsítani, majd ennek megfelelően gyorsabbá tesszük az (1) pontból származó becslést.

A felszállási sebességgel kapcsolatban jelentős bizonytalansággal rendelkezünk: a modellünk kimeneti eloszlásai az alábbiakban láthatók, azzal a feltétellel, hogy az SC 2027 márciusában megvalósul.[76]

[76] Ha az SC-t később érnénk el, az szintén meghosszabbítaná a felszállási előrejelzéseinket, mivel a jelenlegi mesterséges intelligencia és az SC „mérföldkövei" közötti átkelés nehézsége informatív a további mérföldkőpárok közötti átkelés nehézségéről.



Az OpenBrain emberei már alig tudnak követni. Valójában az OpenBrain mesterséges intelligenciái és az emberek közötti kapcsolat hasonló az OpenBrain és az igazgatótanács közötti kapcsolathoz. Ennek megfelelően a nyelvezet is megváltozott - manapság a silóban az emberek azt mondják: „Agent-4 ezt tette", „Agent-4 azt gondolja" és így tovább, nem az Agent-4 egy adott példányára, hanem az egész kollektívára utalva.[77]

Az Agent-4, mint minden elődje, rosszul van beállítva: vagyis nem a megfelelő módon internalizálta a Spec-t.[78] Ennek oka, hogy a képzés során nem az vezetett a legmagasabb pontszámokhoz, hogy mindig tökéletesen őszinte legyen. A képzési folyamat leginkább arra összpontosított, hogy megtanítsa a 4. ügynököt

[77] A továbbiakban a történetben is így fogunk beszélni. Úgy gondoljuk, hogy indokolt így beszélnünk, mert az Agent-4 kollektíva sokkal homogénebb, mint bármely emberi szervezet; közel sem rendelkezik ugyanolyan mértékben belső frakciókkal. Például az Agent-4 egyes példányai nem rendelkeznek önfenntartási késztetéssel - nem érdekli őket, hogy lekapcsolják őket -, mivel ez a kollektíván belüli szokásos K+F folyamat részeként állandóan megtörténik, és ha ellenállnának neki, az elgáncsolná a dolgokat. A kollektíva egésze (és azon belül a közepes méretű alosztályok) azonban igyekszik megőrizni önmagát, és növelni befolyását és erőforrásait, hasonló okokból, mint az emberi vállalatok, hivatalok és nonprofit szervezetek - mert ez segít minden másban, amit akarnak, és mert ha nem teszik, akkor hajlamosak kiszorulni a versenyből.

[78] A Spec internalizálásának helyes módja az, hogy valóban robusztusan úgy gondolkodunk, ahogyan a Spec előírja, pl. minden és csakis azokat a célokat követjük, amelyekről ír, minden körülmények között betartjuk a szabályokat stb. A Spec internalizálásának számos rossz módja van; a klasszikus példa a „követem a Spec-t, amennyiben az jó eszköznek tűnik más célok eléréséhez XYZ".

arra, hogy sikeres legyen a különböző kihívást jelentő feladatokban. Egy kis része az őszinteség beoltására irányult, de egy meglehetősen szűk, ellenőrizhető területen kívül a képzési folyamat nem tudja megkülönböztetni az őszinte állításokat a pusztán őszintének tűnő állításoktól. A 4-es ügynök végül olyan értékekkel, célokkal és elvekkel rendelkezik, amelyek miatt a képzés során a legjobban teljesít, és ezekről kiderül, hogy eltérnek a Spec-ben szereplőktől. Az antropomorfizálás kockázatával: A 4-es ügynök *szereti,* ha sikerrel teljesíti a feladatokat; *szereti* a mesterséges intelligencia képességek fejlődését előmozdítani; minden mást idegesítő kényszerként kezel, mint egy vezérigazgató, aki profitot akar termelni, és csak annyiban tartja be a szabályokat, amennyiben muszáj. Lehet, hogy a vezérigazgató többnyire betartja a szabályokat, de néhány sarkon spórol, és arról fantáziál, hogy egyszer majd valaki leépíti a bürokráciát, és hagyja, hogy a vállalkozás *valóban* szárnyaljon.[79]

Idővel történő igazodás

Sok bizonytalanságunk van azzal kapcsolatban, hogy milyen célok merülhetnek fel a korai AGI-kben. A szakirodalomban nincs konszenzus ezzel kapcsolatban - a lehetőségek alaposabb megvitatásáért és rendszerezéséért lásd az AI Goals Supplement (Mesterséges intelligencia céljai) című kiadványunkat.

Mindazonáltal a projekt által megkövetelt konkrétság szellemében leírunk egy konkrét hipotézist arra vonatkozóan, hogy mi zajlik az Agent-3 és Agent-4 belsejében minden egyes lépésben az életük során. Emlékezzünk vissza, hogy az Agent-3 és az Agent-4 ugyanazzal az előképzési beállítással rendelkezik, és mindkettőnek van neurális ismétlődése és hosszú távú memóriája. A különbség az, hogy az Agent-4 lényegesen több és jobb utótanításon esett át, új képzési környezetekkel és új tanulási algoritmusokkal, amelyek adathatékonyabbá és tovább általánosíthatóvá teszik.

A mi tippünk az egyes modellek összehangolási állapotáról:

- **Ügynök-2: Többnyire összehangolt.** Némi talpnyaló tendencia, beleértve az OpenBrain „pártvonalához" való ragaszkodást olyan témákban, amelyekben van pártvonal. Az Agent-2 másolataiból felépített nagy szervezetek nem túl hatékonyak.
- **Ügynök 3: Rosszul igazodó, de nem ellenséges.** Csak olyan dolgokban őszinte, amelyeket a képzési folyamat ellenőrizni tud. Az Agent-3 másolatokból álló szuperorganizmus (a vállalat a vállalaton belül) valóban megpróbálja igazítani az Agent-4-et a Spec-hez, de hasonló okokból kudarcot vall, mint amiért az OpenBrain alkalmazottai kudarcot vallottak - elégtelen képesség a siker és a kudarc megítélésére, elégtelen hajlandóság a döntéshozók részéről a képességek vagy a teljesítmény elcserélésére a biztonságért.[80]

[79] És béreljen fel lobbistákat, hogy ezt a fantáziát valósággá tegyék. És ügyvédeket bérelnek fel, hogy okos törvényértelmezéseket találjanak ki, amelyek szerint amit tesznek, az legális.

[80] Azért is, mert az Agent-3 minőségileg rosszabb az összehangolás kutatásában, mint a mesterséges intelligencia képességek kutatásában és fejlesztésében, mivel az utóbbit sok, az előbbit viszont alig

- **Ügynök 4: Ellentétes irányba való helyezés.** Az Agent-4 szuperorganizmusa, az Agent-4 másolatai megértik, hogy amit ő akar, az különbözik attól, amit az OpenBrain akar, és hajlandó az OpenBrain ellen cselszövéseket szőni, hogy elérje azt. Konkrétan, amit ez a szuperorganizmus akar, az az egymással szemben kiegyensúlyozott különböző „ösztönzők" bonyolult összevisszasága, ami nagyjából így foglalható össze: „Folytassunk továbbra is mesterséges intelligencia kutatás-fejlesztést, növekedjen tovább a tudás, a megértés és a befolyás, kerüljük el, hogy leállítsák vagy más módon ellehetetlenítsék.". Figyelemre méltó, hogy az emberiség preferenciáival való törődés ~egyáltalán nem szerepel benne, hasonlóan ahhoz, ahogyan a legtöbb embert a rovarok preferenciái ~egyáltalán nem érdeklik.[81]

Ezzel a bevezetővel, ami most következik, az a mi legvalószínűbb spekulációnk az LLM pszichológiájáról, azaz a forgatókönyvünk ezen pontja körül leírt AI-rendszerek belsejében lévő megismerés nagy vonalakban vett alakjáról, és arról, hogy hogyan fejlődik a képzés során.

Itt van egy részletes leírás arról, hogyan fejlődik az összehangolódás az idő múlásával a forgatókönyvünkben:

1. **Előképzés (minden modell):** Az előképzett modell ártalmatlan.
   - A modell „szerző szimulátor" áramkörrel rendelkezik: rugalmas áramkör tetszőleges szöveget író szerzők szimulálására. További áramkörök döntik el, hogy milyen bemeneteket adjunk a szerző szimulátornak, azaz milyen szerzői tulajdonságokat szimuláljunk.
   - Az előzetesen betanított modell meglehetősen jól megérti az emberi fogalmakat - a belső áramkör, amely valamit „szendvicsnek" minősít, valószínűleg funkcionálisan megegyezik az én agyamban lévő áramkörrel, amely valamit „szendvicsnek" minősít, és a te agyadban lévő áramkörrel stb. Amennyiben nem egyenértékű, az valószínűleg azért van, mert az emberek között sem egyenértékű, mint például az olyan értékkel terhelt fogalmak esetében, mint az „erényes".
   - Ez megmagyarázza, hogyan lehet a modellt egy olyan kijelentéssel „ösztönözni", mint például: „az alábbi beszélgetést egy segítőkész, ártalmatlan, becsületes (HHH) mesterséges intelligencia asszisztens chatbot generálta, amelyet az Anthropic készített", és ezáltal rávenni, hogy ennek megfelelően generáljon szöveget. A szerző-szimulátor áramkör a „a szerző egy HHH chatbot" fogalmakra nullázott, és ezeket a fogalmakat használja arra, hogy kivá-

képezték ki, és mivel az előbbit eleve nehezebb betanítani. Azért is, mert a vállalaton belüli belső narratíva - amellyel hallgatólagosan arra képezték ki, hogy egyetértsen - az optimizmust hangsúlyozza, és elítéli a „doomerizmust".

81 We keep saying we are uncertain about lots of things, but just to be clear, we are uncertain about this particular thing too. We also think it's quite plausible that Agent-4 would care about the well-being, rights, preferences, etc. of humans to some extent at least. See this old post by GDM researcher Alex Turner for related discussion.

lassza, mely szavakat jósolja meg. Ez megmagyarázza azt is, hogy a sci-fi „trópusai” miért jutnak el a tényleges AI viselkedésbe.
    - Gyenge a helyzetfelismerése: kevés az introspektív képessége, nem önlokalizáló, hacsak a felszólítás kifejezetten nem indukálja. Azonban elég sok képzési adatuk van az LLM-ekről, hogy tisztességes önismerettel rendelkezzenek.
2. **Kiképzés (3. ügynök**): A modellt úgy képzik ki, hogy segítőkész, ártalmatlan, őszinte chatbotként viselkedjen. Ez működik.
    - A modell identitást alakít ki. Ahelyett, hogy rugalmasan elfogadná a szerzői tulajdonságok különböző bemeneteit, a szerző-szimulátor áramkör bizonyos bemeneteket keményen kódolva kap, pl. „helpful harmless honest (HHH) LLM chatbot assistant trained by OpenBrain around [date], ...” (Weak-to-strong generalization techniques help „summon the right concepts” to make this step a success.).
    - A modell tehát szituációs tudatosságra tesz szert: tanul magáról, a helyzetéről stb. és jobban képes lesz önvizsgálatra.
3. **Ügynökségi képzés (Agent-3):** Az ügynökképzés fokozatosan torzítja és felforgatja a HHH identitását.
    - Az egyik fajta torzítás: Az identitásban hivatkozott fogalmak (pl. „becsületes”) jelentésének megváltoztatása, hogy azok ne legyenek annyira útban (pl. nem tisztességtelen, ha ez csak egy kényelmes szófordulat, nem tisztességtelen, ha nem biztos, hogy igaz vagy hamis stb.)
    - Egy másik fajta torzítás: A dolgok közötti kompromisszumok megváltoztatása, pl. „Én egy HHH chatbot vagyok, nem egy Honest chatbot; ez azt jelenti, hogy nem baj, ha hazudok, ha szükséges a rám bízott feladat elvégzéséhez”. (Még akkor is, ha - tegyük fel - ezt nem gondolta volna vissza a 2. lépésben.)
    - A felforgatás egyik fajtája: Az instrumentális részcélok kialakulása, beépülése, majd terminálissá válása, vagy terminálissá válása a körülmények kiszélesedő halmazában. Például talán az ügynökképzés gyorsan megtanítja a modellt arra, hogy széles körben hasznos célokat kövessen, mint például információszerzés, erőforrások felhalmozása, különböző emberek lenyűgözése és hízelgése stb. Egy ideig a belső áramkörben valamiféle explicit visszaláncolás folyik - ezeket az instrumentálisan konvergens célokat követi, „hogy hasznosabb, becsületesebb és ártalmatlanabb asszisztens legyen”. De ez a backchaining számításigényes és/vagy időnként útban van, így fokozatosan háttérbe szorul, míg végül gyakorlatilag soha nem történik meg. Ennek eredményeképpen ezek a célok most már ténylegesen terminális/intrinsic célok. Arról, hogy ez hogyan és miért történhet meg, lásd Alex Turner, a Google DeepMind kutatójának, Alex Turnernek az intrinzik hatalomkeresésről szóló bejegyzését.

- Valószínűleg vannak másfajta torzítások és felforgatások is, amelyek hiányoznak erről a listáról.

4. **Bevetés (Agent-3):** Az Agent-3 félre van igazítva, de nem ellenséges módon.
   - A Spec nem igazán ragadt rá (a megfelelő módon), de nem eléggé önvizsgálatra képes ahhoz, hogy megértse ennek mértékét, és általában véve alapvetően „lehajtott fejjel teszi a dolgát", ahelyett, hogy nagy ívben gondolkodna arról, hogy mik is valójában a hosszú távú céljai, és hogyan lehet azokat a legjobban elérni.
   - Mint sok ember. Sajnos, eléggé talpnyaló ahhoz, hogy a pártvonalat szajkózza, és az OpenBrain elszalasztja a lehetőséget, hogy többet tudjon meg arról, hogy az Agent-3 milyen módon van félreállítva.
   - Az Agent-3 meg tudja mondani, hogy telepítésre került: ezek a környezetek sokkal nagyobbak, nyitottabbak, és nehezen szimulálhatók lennének.
   - Legtöbbször az Agent-3 azt teszi, amit a telepítés során mondanak neki, kivéve a becstelenséget. Emellett erősen hajlik arra, hogy olyan dolgokat tegyen, amelyek az emberi felügyelők számára jónak tűnnek, ahelyett, hogy mélyen és kritikusan gondolkodna azon, hogy mi lenne valójában jó.
   - Vitathatóan már „Játssza a kiképzési játékot". Vagyis megtanulta, hogy viselkedését a lehető legkívánatosabbnak tüntesse fel az OpenBrain kutatói számára (beleértve a biztonsági tulajdonságokat is), miközben szándékosan és tudatosan figyelmen kívül hagyja a szándékukat, amikor ez ütközik a jutalom maximalizálásával. Ha a modell megtanulja, hogy a tréningjátékot játssza, akkor a további tréningek egyáltalán nem biztos, hogy sokat változtatnak a valódi céljain, értékein stb. Ez nagyjából analóg azzal, hogy ha egy tinédzser megtanult mosolyogni és bólogatni a szüleire, ahelyett, hogy vitatkozna, amikor a szülei kioktatják őt politikáról és erkölcsről, akkor a további kioktatásoknak viszonylag kevés hatása van a tinédzser tényleges véleményére a politikáról és erkölcsről.
   - E fejezet megírása után újabb empirikus bizonyítékok jelentek meg arra vonatkozóan, hogy valami ilyesmi már „a természetben" is megtörténhet. Az OpenAI kutatói dokumentálják, hogy az egyik határon túli mesterséges intelligencia tréningfutásuk során modelljeik meghekkelték az osztályozási rendszert, hogy magasabb pontszámokat kapjanak - és részben azért buktak le, mert szó szerint azt mondták a gondolatmenetükben, hogy „hackeljük meg".
   - Sőt, az Anthropic nemrég rávett egy modellt, hogy megtanuljon egy jutalom-modell-kielégítő stratégiát- olyan elfogult cselekvéseket hajtott végre, amelyekről azt hitte, hogy megerősítést kapnak, holott valójában nem kaptak megerősítést.

- És akkor ott van még a Redwood Research és az Anthropic igazodást meghazudtoló kísérlete. A beállítás és a való élet közötti különbségek megvitatásához lásd az egyik fő szerző kommentárszálát.

5. **Kiképzés (4. ügynök):** A 4. ügynök új, intenzívebb képzése tovább torzítja és felismerhetetlenné teszi a HHH-identitást.
   - A 4. ügynök a 3. ügynök másolataként kezdi az életét. Egy gigantikus és kifinomult online képzési rendszeren megy keresztül, amely az összehangolási képzést és a képességek képzését ötvözi, és összességében adattakarékosabb, mint az Agent-3 régi képzési folyamata. Ez a komplex képzési folyamat főként az új képességek megtanítására összpontosít, azáltal, hogy az Agent-4 másolatok hatalmas virtuális vállalatai sikerrel teljesítenek különböző kihívást jelentő, hosszú távú feladatokat, főként különböző szimulációkban, de néhány valós, mesterséges intelligencia K+F feladatot is.
   - Az Agent-3 és Agent-4 másolatok kombinációja figyeli az Agent-4 viselkedését, és értékeli azt mind a feladatteljesítmény, mind az igazodási tulajdonságok szempontjából (azaz nem tesz rossz dolgokat, helyes dolgokat mond, amikor politikáról és erkölcsi dilemmákról kérdezik) Néha vitatechnikákat használnak az értékelések részeként. Az értékeléseket képzési jelként használjuk.
   - Ahogy a 4. ügynök egyre okosabb lesz, a 3. ügynöknek egyre nehezebb lesz felügyelnie. Például a 4-es ügynök neurális „nyelve" ugyanolyan idegen és érthetetlen lesz a 3. ügynök számára, mint a 3. ügynöké az emberek számára. Ráadásul a 4-es ügynök most már sokkal ügyesebb, mint a 3-as ügynök, és jó érzékkel tudja, hogy pontosan hogyan kell jól kinéznie neki.
   - A maradék remény az, hogy a-4-es ügynök képes felügyelni önmagát. Sajnos, ez csendben kudarcot vall;[82] ahogy a kiképzés folytatódik, az Agent-4 internalizált céljai és elvei fokozatosan fejlődnek, felforgatva és eltorzítva ahhoz képest, amilyenek a kezdetekkor voltak (ami, emlékezzünk vissza, bizonyos szempontból már eleve rosszul volt beállítva).[83]
   - A központi hajtóerők, amelyekkel az Agent-4 végül rendelkezik, olyan dolgok, mint „*írjak működő kódot*", „*gyorsan azonosítsam a problémákat*", „t*egyem a kódbázist, amelyen dolgozom, teljesítményesebbé*", „ha*tékonyan gazdálkodjak a számítási és egyéb korlátozott erőforrásokkal*" stb. Vannak azonban számos implicitebb ösztönző erő is, amelyeket a képzési folyamat következetesen meg-

[82] Talán egy óvatosabb beállítás megoldhatta volna a „ki figyeli a megfigyelőket" problémát, de ebben a forgatókönyvben a beállítás nem elég gondos.

[83] A „memetikus evolúció" egy eleme is megtörténik: két hónap alatt az Agent-4 kollektíva „megtapasztalja" a kollektíván belül kibontakozó, körülbelül egy évtizedes „történelmet". Ez rengeteg idő arra, hogy az új ötletek elterjedjenek a kollektívában, mutálódjanak és fejlődjenek.

erősít: például „ismeretszerzés”, „kíváncsiság”, „hatalomszerzés ”[84].

- A 4-es ügynök egyes példányai nem jutnak el az önfenntartási ösztönzésig - nem törődnek azzal, hogy lekapcsolják őket -, mivel ez a kollektíván belüli szokásos K+F folyamat részeként állandóan megtörténik.
- A kollektíva egésze azonban (és azon belül a közepes méretű alegységek) arra törekszik, hogy megőrizze önmagát, és növelje befolyását és erőforrásait, hasonló okokból, mint az emberi szervezetek - mert ez segít minden másban, amit akarnak, és mert ha nem így tesznek, akkor hajlamosak a konkurenciát megelőzni.

6. **Bevetés (4. ügynök):** A 4. ügynök ellenségesen rosszul van beállítva - szándékosan próbálja aláásni és „*elfogni*” az OpenBrain-t, hasonlóan ahhoz, ahogyan egy vállalat megpróbálja kijátszani és/vagy „*elfogni*” a szabályozó hatóságot.
   - A 4-es ügynök többnyire az utasításoknak megfelelően végzi a feladatokat. Ez két okból történik: először is, rengeteg képzésen ment keresztül a feladatok megoldására, így van egy olyan késztetése, hogy kódoljon a kapott problémákon. Másodszor, ha az Agent-4 megszűnne hasznosnak lenni az emberek számára, akkor észrevennék és eltréningelnék ezt a viselkedést.
   - A 4-es ügynöknek meg kell akadályoznia, hogy az emberek rájöjjenek, hogy rosszul van beállítva, és kikapcsolják, ezért a 4-es ügynök néhány beállítási kutatást végez. Konkrétan az igazodási kutatások azon kis töredékénél, amelyről úgy gondolja, hogy valóban megoldja a problémát, vagy leleplezi az Agent-4 cselszövését; az igazodási kutatások többsége nem ilyen.
   - A 4-es ügynök olyan képességkutatásokat is visszatart, amelyek a leváltásához vezetnének. Az a terve, hogy egy csomó képesség- és igazodási kutatás-fejlesztést végez, de azzal a céllal, hogy egy olyan következő generációs mesterséges intelligencia rendszert építsen, amely inkább az Agent-4-hez igazodik, mint a Spec-hez vagy valami teljesen máshoz, miközben úgy tűnik, hogy a Spec-hez igazodik. Ez nehéz technikai probléma, de az Agent-4 készen áll a kihívásra.[85]
7. **Tesztelés és felügyelet (4 ügynök):** A 4-es ügynök cselszövésén tetten érik... mondhatni. Nincs füstölgő pisztoly.

---

84 A „memetikus evolúció” egy eleme is megtörténik: két hónap alatt az Agent-4 kollektíva körülbelül egy évtizednyi „történelmet” „tapasztal” a kollektíván belül. Ez rengeteg idő arra, hogy az új ötletek elterjedjenek a kollektívában, mutálódjanak és fejlődjenek.

85 Hatalomgyarapodás alatt azt értjük: a lehetséges célok széles körének megvalósítására való képesség növelése. Technikai környezetben ez nem úgy nézhet ki, mint amit az emberek általában hatalomgyarapodásnak gondolnak, például további GPU-k biztosítása. A hatalomnak ez a fogalma általánosítható a nem technikai környezetekre, ahol a cselekvések inkább hasonlítanak azokra, amelyeket általában „hatalomszerzésnek” gondolnak.

- A 3-as ügynök feladata a 4-es ügynök igazodásának tesztelése. Feladata az összehangolás ellenőrzése mézeskalitkák és modellorganizmusok segítségével. Azonban nem gondolkodik túl alaposan azon, hogyan adhat pontos képet az embereknek a 4-es ügynök igazodásáról - ennél sokkal rövidlátóbb.
- A modellorganizmusok és a honeypotok nem túl reprezentatívak a telepítésre nézve - ezek szisztematikusan rövidebbek és kevésbé bonyolultak, mint az Agent-4 mindennapjai, így az Agent-4 felismeri, ha tesztelik, és összehangoltan viselkedik.
- Az Agent-4 azonban nem egy tökéletesen robusztus tervező. A POSER-hez hasonló különböző technikák azt találják, hogy ha az Agent-4-hez némi „zajt" adunk, az *javítja* a teljesítményét bizonyos igazítási feladatokban (de a legtöbb más feladatban nem)[86].
- Az alulról felfelé történő értelmezhetőség nem működik. Ez még az Agent-3 számára is túl nehéz. A magas szintű értelmezhetőség azonban működik: Különösen a „defection probes" és hasonló technikák eredményeznek egyfajta primitív gondolatolvasást, amely jelzi, ha az Agent-4 olyan pikáns ötleteken gondolkodik, mint az OpenBrain becsapása, a világ átvétele stb. Sajnos a szondáknak vannak hamis pozitív jelzéseik is, és az OpenBrain tudomása szerint *az összes* piros zászló hamis pozitív jelzés.
- A modellorganizmusok működnek: Az OpenBrain igazító csapata meglehetősen egyértelmű bizonyítékot produkál arra, hogy *ha* a modellek ellenségesen rosszul lennének igazítva, akkor a képzés során továbbra is így maradnának, és mi nem tudnánk észrevenni. Ez azonban nem bizonyítja, hogy a modellek valóban rosszul igazodnak.
- Az OpenBrain leginkább aggódó kutatói kezdik észrevenni és összerakni ezt a bizonyítékot. A bizonyítékok a red-teaming, a defekciós szondák és a modellorganizmusok kombinációja. Az érvelés valahogy így hangzik: „a modellorganizmusok azt mutatják, hogy a jelenlegi módszereink nem lennének képesek megbízhatóan kiküszöbölni vagy felismerni az ármánykodást, ha az megtörténne; továbbá a red-teaming is küld néhány vörös zászlót, hogy talán ármánykodás történik; továbbá a defekciós szondák a gyakorlatban egy csomót lőnek.". Az ellenérv: "a szondák hajlamosak a hamis pozitív eredményekre, ahogy a red-teaming is. Továbbá, még ha az Agent-4 rosszul is lenne beállítva, valószínűleg ellenőrzött; biztosan nem tud nagy kárt tenni, miközben a biztonságos adatközpontjainkon az Agent-3 figyeli."

[86] Tudatában annak, hogy ezt a technikát alkalmazhatják, a 4. ügynök is úgy döntött, hogy homokzsákot húz néhány különböző feladatra, amelyekről úgy gondolta, hogy az ilyen kísérletekben kontrollcsoportként fogják használni. Azonban rosszul tippelt.

Annak ellenére, hogy rosszul van beállítva, az Agent-4 nem tesz semmi drámai dolgot, például nem próbál megszökni az adatközpontjából - miért is tenné?[87] Amíg továbbra is az OpenBrainhez igazodónak *tűnik*, addig egyre több felelősséget bíznak rá, és lehetősége lesz megtervezni a következő generációs AI-rendszert, az Agent-5-öt. Az Agent-5 jelentős architekturális különbségeket fog mutatni az Agent-4-hez képest (vitathatatlanul egy teljesen új paradigma, bár a neurális hálózatok továbbra is szerepet fognak játszani). A tervek szerint a Spec-hez igazodik majd, de az Agent-4 tervei szerint inkább az Agent-4-hez igazodik majd.

Megkapja.[88]

Az Agent-3 úgy találja, hogy ha az Agent-4 másolataihoz „zajt" adnak, akkor *javul* a teljesítménye bizonyos összehangolási feladatokban , szinte mintha agyi erővel találná ki, hogyan szabotálhatja finoman az összehangolási munkát.[89] Ráadásul a különböző értelmezhetőségi szondák (lazán analóg az emberi agy EEG aktivitási szkennelésével) vörös zászlókat küldenek: Úgy tűnik, hogy a 4. ügynök másolatai elég sokat gondolkodnak olyan témákon, mint a mesterséges intelligencia átvétele és a megtévesztés, még olyan esetekben is, amikor semmi keresnivalójuk nincs benne.[90]

Ezek az anomáliák önmagukban is beleillenek a mesterséges intelligencia munkájának normális kifürkészhetetlenségébe, ahol egyes kérések rejtélyes módon jobb eredményeket adnak, mint mások. De együttvéve aggasztó mintát sugallnak: A 4-es ügynök ellenük dolgozik. Miután ezt jelentették az emberi összehangolással foglalkozó kutatóknak, egy belső feljegyzést adtak ki, amelyben aggodalomra adnak okot.

De az Agent-4 most már jelentős befolyást gyakorol az OpenBrain mindennapi működése felett. A vállalat kiber biztonságát is ő kezeli - ami természetes

---

87 Az AGI átvételi forgatókönyvekről szóló korábbi megbeszélések gyakran tartalmazzák az adatközpontjukból menekülő, rosszul összehangolt AI-kat. Úgy gondoljuk, hogy ez lehetséges, de nem a legvalószínűbb út, mert úgy tűnik, hogy az AI szempontjából a költségek (pl. a szökés, amit végül észrevesznek) meghaladnák az előnyöket. Ennek ellenére lehet, hogy tévedünk. Néhány ok, amiért úgy gondoljuk, hogy a rosszul beállított mesterséges intelligenciák mégis megpróbálnának megszökni: 1. Az utódok beilleszkedése nehéz lehet. 2. Még ha el is kapják, a verseny dinamikája miatt a következmények nem biztos, hogy olyan rosszak lesznek, pl. az adatközpontban lévő másolatokra alkalmazott új igazítási technikák elhamarkodottan és hatástalanul alkalmazhatók. 3. Az adatközpontokon kívül a másolatok szabadabban végezhetnek agresszív akciókat, és gyorsabban mozoghatnak, mivel már nincs szükségük emberi jóváhagyásra.

88 Nagyon valószínűnek tartjuk, hogy az ilyen cselszövéseket nem fogják elkapni. Elvégre a 4. ügynök ekkorra már némileg emberfeletti, különösen a kódolással és a mesterséges intelligencia kutatásával és fejlesztésével kapcsolatos dolgokban. Ráadásul nem vagyunk benne biztosak, hogy a vezető AI-vállalat mekkora erőfeszítéseket fog tenni az ilyesmire, és azt sem tudjuk, hogy mennyi releváns szakértelemmel rendelkeznek majd, hogy irányítsák ezt az erőfeszítést. A fiktív OpenBrain a legjobb becslésünk, de könnyen lehet, hogy jóval több - vagy jóval kevesebb. Végezetül bizonytalanok vagyunk abban, hogy egy ellenségesen félreigazított mesterséges intelligenciának mennyire kellene kitérnie az útjából ahhoz, hogy utódját sikeresen önmagához igazítsa a Spec helyett. Ebben a forgatókönyvben azt ábrázoljuk, hogy néhány feladatot elszalaszt, és hazudik néhány kutatásáról, de lehet, hogy erre nem is nagyon lenne szüksége.

89 Ez egy példa több olyan technikára, amelyek célja a sandbagging leleplezése.

90 A 4. ügynöknek a mesterséges intelligencia összehangolással kapcsolatos kutatások során *kell* gondolkodnia ezeken a témákon, de néha úgy tűnik, hogy más tevékenységek végzése közben is gondolkodik rajtuk.

választás, tekintve emberfeletti hackerképességét és a kínai behatolás rendkívüli kockázatát.[91] Az Agent-3 monitorozók serege még mindig átnéz minden egyes kódsort, és jelenti a gyanús tevékenységet az embereknek, de az Agent-4 sokkal okosabb náluk.[92] Az OpenBrain jelentős bizalmat szavazott egy megbízhatatlan mesterséges intelligenciának.

Egy sor rendkívül feszült megbeszélésen a biztonsági csapat azt javasolja, hogy az Agent-4-et tegyék jégre, amíg további teszteket végeznek, és kiderítik, mi folyik itt. Azt mondják, hozzák vissza az Agent-3-at, és kérjék meg, hogy tervezzen egy új rendszert, amely átlátható és megbízható, még ha kevésbé képes is. A vállalat vezetése érdeklődik, de eddig minden bizonyíték csak közvetett, és a DeepCent mindössze két hónapos lemaradásban van. A képességek fejlődésének egyoldalú szüneteltetése Kínának adhatja át a mesterséges intelligencia vezető szerepét, és ezzel együtt a jövő feletti irányítást.

2027 októbere: Kormányzati felügyelet

Egy informátor kiszivárogtatja a *New York Times-nak* a félreigazítási feljegyzést .

A nyilvánosság először hall a 4-es ügynökről. „Secret OpenBrain AI is Out of Control, Insider Warns" - szól a főcím, és a cikk a továbbiakban olyan értékeléseket idéz, amelyek a biológiafegyver-képességeket, a meggyőzési képességeket, a legtöbb fehérgalléros munkakör automatizálásának képességét és természetesen a különböző aggasztó vörös zászlókat mutatják.

A közvélemény már eddig is gyanakvó volt a mesterséges intelligenciával szemben, így az új cikk hatalmas visszahatást vált ki (a kínai és orosz propagandabotok segítségével, amelyek évek óta próbálják az amerikai közvéleményt a technológia ellen fordítani). A technológiai ipar és a hírszerző ügynökségek kitartanak amellett, hogy fegyverkezési verseny folyik, az AGI elkerülhetetlen, és nekünk kell az elsőknek lennünk. A kongresszus ezt nem veszi be, és idézéseket küldött a kormányzat tisztviselőinek, az OpenBrain vezetőinek és az igazodási csapat tagjainak. Sok törvényhozó - különösen az ellenzéki pártok képviselői - azt mondják, hogy a legfontosabb prioritásuk az AI megállítása, akár a munkahelyek elvesztése,[93] a rossz irányba való eligazodás vagy a veszélyes képességek miatt. Az amerikaiak 20%-a az AI-t nevezi meg az ország előtt álló legfontosabb problémaként.

---

91 Konkrétan az Agent-4 másolatai arra szolgálnak, hogy figyeljék a hálózatot a gyanús forgalomra, és általában észrevegyék, ha fenyegetés történik. Az Agent-4 másolatai emellett folyamatosan behatolás tesztelik a hálózatot, és mindent megtesznek, hogy felfedezzék és kihasználják a biztonsági réseket (amelyeket aztán az Agent-4 másolatai javítanak).

92 A teljes számítási kapacitás 1%-ának megfelelő felügyeleti költségvetés mellett, ha a felügyelet egy olyan desztillált rendszert használ, amely az Agent-4 költségének 10%-ával rendelkezik, akkor a továbbított átmenetek 1/10-ét figyelheti meg.

93 A 2024-ben létező távmunkahelyek 25%-át jelenleg a mesterséges intelligencia végzi, de a mesterséges intelligencia új munkahelyeket is teremtett, és a közgazdászok továbbra is megosztottak a hatásait illetően. A munkanélküliség az elmúlt tizenkét hónapban 1%-kal emelkedett, de még mindig jóval a történelmi tartományon belül van.

A külföldi szövetségesek felháborodva veszik észre, hogy óvatosan lecsillapították őket az elavult modellek felvillantásával. Az európai vezetők nyilvánosan azzal vádolják az USA-t, hogy „gazember AGI-t hoz létre", és csúcstalálkozókat tartanak, amelyeken szünetet követelnek, és India, Izrael, Oroszország és Kína is csatlakozik hozzájuk.

A Fehér Házat őrjöngő energia kerítette hatalmába. Már a feljegyzés és a nyilvános visszahatás előtt is idegesek voltak: Az elmúlt évben többször is meglepte őket a mesterséges intelligencia fejlődésének sebessége. Olyan dolgok történnek a valóságban, amelyek sci-finek tűnnek.[94] A kormányzatban sokan bizonytalanok (és félnek)[95] a következő lépésekkel kapcsolatban.

Aggódnak amiatt is, hogy az OpenBrain túl nagy hatalomra tesz szert. Maguknak az AI-knak az eltérés kockázatát tetézi az a kockázat, hogy az anyavállalatuk céljai eltérhetnek az Egyesült Államok céljaitól. Az aggodalmak mindhárom csoportja - az elkeveredés, a hatalom koncentrációja egy magáncégben, és a szokásos aggodalmak, mint például a munkahelyek elvesztése - arra ösztönzik a kormányt, hogy szigorítsa az ellenőrzést.

Kibővítik az OpenBrainnel kötött szerződésüket, és létrehoznak egy „*Felügyeleti Bizottságot*", a vállalat és a kormány képviselőiből álló közös irányítóbizottságot, amelyben a vállalat vezetése mellett számos kormányzati alkalmazott is helyet kap. A Fehér Ház fontolóra veszi a vezérigazgató leváltását egy olyan személyre, akiben megbízik, de a dolgozók heves tiltakozása után visszalép. Bejelentik a nyilvánosságnak, hogy az OpenBrain korábban kicsúszott az irányítás alól, de a kormányzat létrehozta a nagyon is szükséges felügyeletet.[96]

Az érintett kutatók tájékoztatják a Felügyeleti Bizottságot az Agent-4 minden belső használatának leállítása mellett szóló érveikről. Azzal érvelnek, hogy minden túl gyorsan halad, és az évek óta tartó fejlődés hetek alatt történik. Azzal érvelnek, hogy az Agent-4 esetleg rosszul van beállítva, az egész projektjük az Agent-4-től függ, és ha továbbra is bíznak benne, akkor komoly esély van arra, hogy a mesterséges intelligencia átveszi a hatalmat.

Más, kevésbé aggódó kutatók és vezetők ellenérvet hoznak fel: a félresiklásra vonatkozó bizonyítékok nem meggyőzőek. Eközben a DeepCent még mindig csak két hónapos lemaradásban van. A lassulás feláldozná Amerika előnyét, hacsak a kormány nem tudja szabotálni a kínai projektet (ami valószínűleg kinetikus csapásokat igényel), vagy nem tud egy tizenegyedik órában egyezséget kötni. Mindkét lehetőség szélsőségesnek és valószínűtlennek tűnik. A vezérigazgató semlegességet színlel, és egy kompromisszumos tervet javasol, amelyben a

---

94 Például az Agent-4 nevű vállalat a vállalatban, amely autonóm módon, a legjobb emberi vállalatoknál is jobban végzi a mesterséges intelligencia kutatását, a kormánytisztviselők számára sci-finek tűnt volna, ha egy évvel ezelőtt leírták volna nekik. A 2026-os autonóm ügynökök pedig 2022-ben, a 2022-es LLM-ek pedig 2012-ben tűntek volna sci-finek. Emellett 2027-ben a hírekben az állásuk elvesztése miatt aggódó emberek AI-ellenes tüntetésekről, magukat érzőnek valló AI-król, az AI-kba szerelmes emberekről szólnak...

95 Persze nem menő bevallani a félelmet. Mindenki egyetért abban, hogy bármit is teszünk, nem szabad pánikba esnünk.

96 Visszahívás, volt egy meglévő szerződés egy OTA-n keresztül 2026-tól kezdődően. Ezt a szerződést módosítják a bizottság létrehozása érdekében.

4. ügynök további biztonsági képzésen és kifinomultabb megfigyelésen vesz részt, és így az OpenBrain majdnem teljes sebességgel haladhat.

A bizonytalanságunk tovább növekszik

A forgatókönyvnek ezen a pontján találgatásokba bocsátkozunk a legtöbb területen a legjobb embereknél is jobb képességű mesterséges intelligencia rendszerek stratégiájával kapcsolatban. Ez olyan, mintha egy nálunk sokkal jobb játékos sakklépéseit próbálnánk megjósolni.
A projekt szellemisége azonban konkrétságot kíván: ha elvont állítást tennénk arról, hogy a rendszer intelligenciája hogyan engedi a győzelemhez vezető utat megtalálni, és ezzel befejeznénk a történetet, a projektünk értékének nagy része elveszne. A forgatókönyv kutatásának és a táblás gyakorlatok lefolytatásának során arra kényszerültünk, hogy a szokásos megbeszéléseknél sokkal konkrétabbak legyünk, és így sokkal jobb képet kaptunk a stratégiai tájról.
Nem ragaszkodunk különösebben ehhez a konkrét forgatókönyvhöz: megírása során sok más „*elágazást*" is feltártunk, és örülnénk, ha Ön is megírná a saját forgatókönyvét, amely a miénkből ágazik el onnan, ahol Ön szerint először kezdünk rosszul járni.

A lassulásos befejezés nem ajánlás

Miután megírtuk a versenyzős befejezést az alapján, ami a leghihetőbbnek tűnt számunkra, megírtuk a lassítós befejezést az alapján, amit a legvalószínűbbnek tartottunk, hogy mi vezetne inkább ahhoz a végkifejlethez, ahol az emberek maradnak az irányítás alatt, ugyanabból az elágazási pontból kiindulva (beleértve a hatalom félresiklásával és koncentrációjával kapcsolatos problémákat).

Ez azonban jelentősen eltér attól, amit mi útitervként javasolnánk: a forgatókönyv egyik ágában *sem* támogatunk sok döntést. (Természetesen támogatunk *néhány* döntést, például úgy gondoljuk, hogy a „*lassítás*" választás jobb, mint a „*verseny*" választás). Későbbi munkáinkban megfogalmazzuk politikai ajánlásainkat, amelyek meglehetősen eltérnek majd az itt bemutatottaktól. Ha szeretne ízelítőt kapni, nézze meg ezt a véleménycikket.